НАЦИОНАЛЬНАЯ АКАДЕМИЯ НАУК УКРАИНЫ
ИНСТИТУТ ПРОБЛЕМ РЕГИСТРАЦИИ ИНФОРМАЦИИ

А.Г. Додонов, Д.В. Ландэ, В.Г. Путятин

КОМПЬЮТЕРНЫЕ СЕТИ И АНАЛИТИЧЕСКИЕ ИССЛЕДОВАНИЯ

Киев 2014

УДК 004.5
ББК 22.18, 32.81, 60.54
К95

Рекомендовано к изданию
Ученым советом Института проблем регистрации информации НАН Украины
(протокол № 3 от 4 ноября 2014)

Рецензенты:
Евдокимов В.Ф. – доктор технических наук, профессор, член-корр. НАН Украины
Качинский А.Б. – доктор технических наук, профессор
Литвинов В.В. – доктор технических наук, профессор

К95 **Додонов А.Г. Компьютерные сети и аналитические исследования** /А.Г. Додонов, Д.В. Ландэ, В.Г. Путятин. – К.: ИПРИ НАН Украины, 2014. – 486 с.

Монография посвящена теоретическим и технологическим основам систем поддержки аналитических исследований в глобальной сетевой среде, методам и средствам мониторинга, агрегирования и обобщения информационных потоков большого объема в компьютерных сетях. Рассматриваются модели и технологии информационного поиска, содержательного анализа текстов и информационных сетей, – базовые понятия в области построения современных аналитических систем.

Книга ориентирована на специалистов в областях автоматизации аналитической деятельности и информационных технологий, студентов старших курсов, аспирантов соответствующих специальностей.

ISBN 978-966-02-7422-8

ИПРИ НАН Украины
Заказ № 9616
Тираж 300 экз.

УДК 004.5
ББК 22.18, 32.81, 60.54

СОДЕРЖАНИЕ

«*Эксперт – это человек, который совершил все возможные ошибки в очень узкой области*»

Нильс Бор, лауреат Нобелевской премии

«*Нам не дано предугадать, Как слово наше отзовется…*»

Ф. Тютчев

ПРЕДИСЛОВИЕ

В настоящее время человечество столкнулось с парадоксальной ситуацией, заключающейся в том, что резко увеличивающиеся объемы информации приводят к снижению общего уровня информированности, многократно усложняют аналитическую деятельность в различных областях.

Сегодня, когда сеть Интернет превратилась во всемирную самодостаточную медиасреду, становится одним из самых важных источников информации, она все чаще не устраивает не только профессионалов-аналитиков, но и обычных пользователей. Это происходит несмотря на то, что практически все известные поисковые сервисы охватывают возможности поиска документов в различных форматах, обращения к новостным сообщениям, полнотекстовым документам, мультимедийным файлам. Сетевая аналитика, включающая анализ информационного наполнения современных компьютерных сетей, охват и обобщение динамических информациооных массивов сверхбольшого объема (Big Data), потоков информации, непрерывно появляющихся в глобальной сетевой среде, требует новых подходов, обеспечивающих выявление наиболее важных фрагментов в информационных потоках, ориентации на конкретных потребителей или на целевые группы.

Для обеспечения информационно-аналитической деятельности в самых различных сферах оказывается

необходимым реализовать теоретически обоснованные информационные технологии, охватывающие все цепочки работы с информацией, включая мониторинг, агрегирование потоков информации, ее аналитическое обобщение, визуальное отображение.

Именно проблематике теоретических основ и информационных технологий поддержки информационно-аналитической деятельности посвящена данная книга, в которой рассматривается широкий спектр вопросов, связанных с развитием информационного пространства, динамикой и содержанием информационных потоков, возможностями навигации и поиска в сетевом информационном пространстве, технологиями современных информационно-поисковых систем, методами аналитической обработки и обобщения информации, технологическими основами мониторинга, адаптивного агрегирования и обобщения информации, и, наконец, практическими вопросами построения адаптированного документального хранилища, ориентированного на задачи информационно-аналитической деятельности.

Несмотря на то, что данная книга в основном ориентирована на специалистов, студентов и аспирантов в области автоматизации процессов обработки аналитической информации, хочется верить, что она будет также полезна и аналитикам-практикам, специалистам в области конкурентной разведки, которые с помощью представленных здесь методологических подходов и информационных технологий смогут повысить эффективность и качество своей работы.

ВВЕДЕНИЕ

Аналитическая деятельность в самых различных областях предполагает работу с информацией, ее глубоким осмыслением, принятием адекватных решений по анализу той или иной ситуации, с получением дополнительной информации, анализом всей имеющейся информации, относящейся к рассматриваемой проблеме, тематическую обработку информации, подготовку и визуализацию аналитических отчетов, их верификацию, получение управленческих решений на базе новых знаний.

В настоящее время развитие методов и средств мониторинга, адаптивного агрегирования и обобщения потоков информации из глобальных компьютерных сетей для поддержки информационно-аналитической деятельности в различных прикладных сферах является весьма актуальной проблемой.

В то же время, своевременное получение многоаспектной и объективной информации по компьютерным сетям для дальнейшего ее использования в разнообразной аналитической деятельности требует применения современных технологических решений.

Эволюция освоения пользователями-аналитиками информационного пространства охватывает три основных этапа.

Начиная работу в Интернете, пользователи сначала обращаются к избранным источникам, – веб сайтам, архивам или базам данных (БД), постоянно отслеживают изменения, динамику появления новых материалов, т.е. самостоятельно проводят мониторинг этих сегментов информационного пространства.

Следующим шагом использования ресурсов веб-пространства обычно является применение сетевых информационно-поисковых систем (ИПС), каждая из которых имеет свои особенности, но «монополистами» среди которых сегодня являются, безусловно, системы Google, Baidu и Яndex (первая – международная, вторая

– для китайских, а последняя для российских и украинских пользователей).

Обычно опытный пользователь также обращается и к специализированным сетевым метапоисковым системам (МПС), которые агрегируют возможности обычных сетевых ИПС. Некоторые из таких систем имеют возможность адаптации под информационные потребности своих пользователей.

Особенностью информационного пространства является наличие многосторонних, многопрофильных источников информации. Совокупность потоков от этих источников образует информационную среду, обеспечивающую как потребление и накопление, так и расширенное воспроизводство информации. В связи с этим первый раздел данной монографии посвящен понятию информационного пространства и информационным потокам. Приведено качественное и формальное определение информационных потоков, которые понимаются как совокупность находящихся в непрерывной динамике постоянно эволюционирующих взаимосвязанных информационных сообщений – документов, которые создаются, развиваются, модифицируются, утилизируются. Рассмотрены различные подходы к изучению информационных потоков – от моделей в виде простейших линейных уравнений – до современных синергетических систем. В первой главе также рассматриваются источники информации в сети Интернет, а также модели и механизмы формирования кластеров близких по содержанию документов, образующих в динамике тематические или событийные информационные сюжеты.

Вторая глава посвящена алгоритмам и методам поиска информации, навигации в глобальных информационных сетях. Большое внимание в этой главе уделено метапоисковым системам, моделям и технологиям децентрализованного поиска, пиринговым системам.

В третьей главе рассматриваются вопросы

содержательного анализа информационных потоков, связанные с концепцией глубинного анализа текстов (Text Mining). В этой главе рассматриваются методы семантической обработки информации, алгоритмы классификации и кластерного анализа, методы экстрагирования понятий, автоматического построения отчетов. Естественно, большое внимание в третьей главе посвящено вопросам компьютерной лингвистики, вопросам дублирования, сходства и упорядочения текстовой информации. Также в этой главе отражена взаимосвязь современной концепции сложных сетей (Complex Network) и задач компьютерной лингвистики.

Четвертая глава посвящена технологическим основам адаптивного агрегирования информационных потоков, т.е. формирования фрагментов информационного пространства, адаптированных под потребности пользователей. Агрегирование информации логически связано с ее разделением на частичные потоки, соответствующие конкретным запросам, формируемым кластерам в информационном пространстве. В этой главе также подробно рассмотрены принципы построения корпоративных метапоисковых систем. В качестве одного из примеров реализации концепции агрегирования информации в четвертой главе рассматривается корпоративная система Doc's Bundle, ориентированная на обработку потоков документов, представленных в формате PDF. Одним из элементов системы Doc's Bundle является метапоисковый блок, реализующий агрегацию документальных потоков – результатов поиска в крупнейших сетевых поисковых системах, таких как Google, Яndex, Bing и т.д. В четвертой главе также детально рассмотрен типовой интерфейс пользователя корпоративной метапоисковой системы.

В настоящее время большое распространение получили так называемые информационные операции, противодействие различных сил в информационном пространстве. Такие противодействия в виртуальном мире обычно сопровождают противодействия в реальном мире, так называемые информационные

войны. Информационные операции базируются на различных информационных влияниях, манипуляциях. Именно эта тематика нашла свое отражение в пятой главе, которая охватывает также такие вопросы, как этапность информационных операций, выявление и анализ, моделирование информационных операций и противодействия им. В этой же главе рассматривается явление так называемой «сетевой мобилизации», как средства объединения усилий участников виртуального пространства социальных сетей для решения некоторых проблем, возникающих (или имеющих место) в реальном мире.

1. ИНФОРМАЦИОННОЕ ПРОСТРАНСТВО

1.1. Понятие информационного пространства

Под информационным пространством (ИП) принято понимать совокупность информационных ресурсов, технологий их сопровождения и использования, информационных и телекоммуникационных систем, образующих информационную инфраструктуру. Элементами информационного пространства могут быть, в частности, документы, обобщающие самые различные виды информации – файлы, электронные письма, веб-страницы независимо от форматов их представления.

Приведенное определение информационного пространства является качественным. Следует отметить, что термин «пространство» в данном случае не совпадает с понятием «пространство» в математике или физике, вместе с тем оно является отражением реальной действительности (жизни), благодатной средой для моделирования общественных процессов.

В качестве примеров удачных моделей ИП можно привести «векторно-пространственную» модель Г. Солтона [Salton, 1975] или модель старения информации Бартона-Кеблера [Bruton, 1960]. Модель такого информационного пространства, как сеть WWW, была построена А. Бредером и его соавторами из компаний IBM и Altavista [Broder, 2000].

В 1999-м г. А. Бредер из компании IBM и его соавторы из компаний AltaVista, IBM и Compaq сделали первую попытку математического описания структуры веб-пространства – «карты» его ресурсов и гиперсвязей, получившей благодаря своей форме название «галстука-бабочки» (Bow Tie). С помощью баз данных и поискового механизма AltaVista было проанализировано свыше 200 млн. веб-страниц и несколько миллиардов ссылок, размещенных на этих страницах [Broder, 2000], [Blekanov, 2014].

В рамках общей задачи определения структуры

связей между отдельными веб-страницами было выявлено центральное ядро (28% веб-страниц) – зона сильной связности сети (Strongly Connected Component, SCC), "отправные веб-страницы" (IN), охватывающие 22% ресурсов, «конечные веб-страницы» (OUT), также охватывающие 22% ресурсов, «отростки, мысы и перешейки» (22% веб-страниц). Существуют и "острова", которые вообще не пересекаются с остальными ресурсами Интернет (рис. 1).

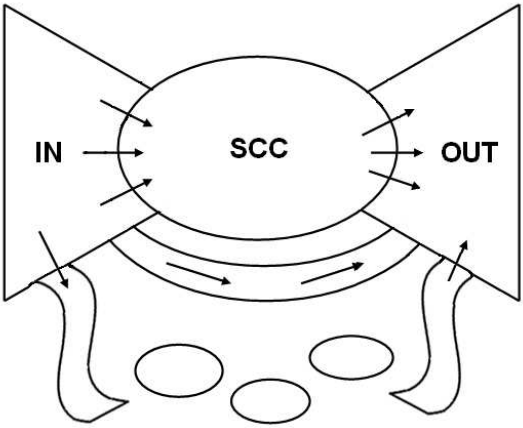


Рис. 1 – Модель веб-пространства Bow Tie

Было обнаружено, что пропорции названных категорий в течение нескольких месяцев оставались неизменными, несмотря на значительное увеличение общего объема веб-ресурсов. Топология и характеристики модели оказались примерно одинаковыми для различных подмножеств веб-пространства, подтверждая тем самым наблюдение о том, что свойства структуры всего веб-пространства Bow Tie также верны и для его отдельных подмножеств. Таким образом, алгоритмы, использующие информацию о структуре веб-пространства, предположительно будут работать и на отдельных его подмножествах.

Оказалось, что распределение степеней узлов (входящих и исходящих гиперссылок) веб-пространства

(исследовались сайты домена edu в количестве 325729) подчиняется степенному закону, т.е. вероятность того, что соответствующая степень вершины равна i, пропорциональна $1/i^k$ (для входящих ссылок $k \approx 2.1$, а для исходящих $k \approx 2.45$). Кроме того, оказалось, что сеть WWW является «малым миром» со средней длиной кратчайшего пути, равной 11 и относительно большим значением коэффициента кластерности, приблизительно равным 0.15 (для классического случайного графа это значение составило бы 0,0002).

Вместе с тем необходимо отметить некоторую некорректность расчета объемов «островов» по Бредеру из-за того, что список веб-ресурсов был получен из базы данных системы AltaVista, полученный в результате работы программы-робота, сканирующего веб-ресурсы, переходя от одного веб-ресурса к другому по гиперссылкам.

Модель Бредера не учитывает особенностей динамической части веб-пространства, формируемой потоками новостных сообщений. Применение модели «галстука бабочки» к динамической составляющей веб-пространства сегодня уже нельзя считать корректным по ряду причин, которые будут рассмотрены ниже.

Понятие «информационное пространство» является системообразующим элементом общества, представляющим собой совокупность информационных ресурсов и инфраструктуры, т.е. всю сферу формирования, распространения и использования социальной информации, имеющую своей целью обеспечение полноценного функционирования других элементов социума в целом. Из этого следует сделать вывод, что понятие «информационное пространство» функциональнее и ближе по смыслу термину «связь с общественностью» или «информационное общество».

Информационное пространство представляет собой среду циркулирования информационных потоков и физические средства ее функционирования, поддержки и развития. В состав технологических и

организационных компонентов ИП в обобщенном варианте входят [Коваленко, 2013]:

- Информационно-телекоммуникационная инфраструктура;
- Информационные ресурсы;
- Методы и средства прикладной математики – алгоритмы и программные средства (комплексы), обеспечивающие функционирование аппаратных платформ (систем);
- Организационные меры, обеспечивающие функционирование компонентов информационного пространства;
- Правовые меры (нормы) – информационное законодательство, международные соглашения и договоры, другие национальные и международные нормативные правовые акты.

Информационное пространство можно рассматривать и как множество связанных по смыслу элементов (документов), образующих информационные системы – кластеры близких по тематике документов. При этом оно за все время своего существования сохраняет свои устойчивые закономерности. Многочисленными исследованиями показано, что параметры частотного и рангового распределений документов во многих информационных системах остаются одинаковыми, и определяются параметрами, зависящими от содержания, тематики информации. В связи с этим С.А. Иванов [Иванов, 2002] заметил, что «информационное пространство – это документальная среда, в которой формируются кластерные структуры научных публикаций в периодических изданиях, являющиеся фракталами».

Информационные системы отражают в информационном пространстве коммуникационные процессы в своей тематической области, появление

новых тематик сопровождается возникновением новых фрактальных массивов в информационном пространстве.

Как и многие другие сложные системы, ИП можно представить как коммуникационную среду – в виде системы с комплексом связей информационных источников и преобразователей между собой, влияющих друг на друга в зависимости от уровня восприятия генерируемых и преобразуемых ими отдельных информационных сообщений. При этом для моделирования источников и преобразователей информации, с одной стороны, вполне подходит классическая теория информации как математическая теория связи, разработанная Шенноном в 40-х годах XX-го столетия и существенно дополненная и расширенная в последующие годы работами Н. Винера, В. А. Котельникова и А. Н. Колмогорова. Однако классическая теория информации не учитывает взаимодействия между источниками и преобразователями информации, что, с другой стороны, вполне укладывается в идеологию современной теории сложных сетей (Complex Networks).

1.2. Информационные потоки в информационном пространстве

1.2.1. Качественное описание информационных потоков

Сетевые структуры в информационном пространстве состоят из отдельных элементов, образующих информационные потоки в динамике своей эволюции (появление, развитие, модификация, уничтожение).

Для исследования современных информационных потоков в Интернет, т. е. большого потока сообщений, которые публикуются на страницах веб-сайтов, в социальных сетях, блогах, и т.п., должен применяться принципиально новый инструментарий, так как классические методы обобщения информационных массивов (классификации, фазового укрупнения,

кластерного анализа и т.д.) не всегда пригодны для адекватного отражения состояния динамической составляющей информационного пространства [Брайчевский, 2005]. В этом случае речь идет не столько об анализе документальных массивов фиксированных размеров, пусть даже очень больших, сколько об обобщении динамического потока гипертекстовых данных.

Конечно, большая часть информации, которая представлена в Интернет, находит своего потребителя. Однако, если рассматривать всю совокупность сетевых публикаций как какую-то общность по отношению к конкретному пользователю (или группы пользователей), то можно увидеть ряд проблем, связанных с полнотой, релевантностью и оперативностью получения данных. Поиск, фильтрация, сбор информации в сети Интернет требуют достаточной квалификации персонала и, к сожалению, при этом не могут учитываться все особенности информационной структуры сети и представления в ней данных. Это, в свою очередь, ведет к тому, что единичные выборки информации из веб-пространства не могут считаться репрезентативными.

При этом информационный поток, который «потребляется» конкретным пользователем, носит, как правило, выраженную предметную направленность, которая характеризуется областью его интересов. Поиск и обработка информации в ручном режиме – достаточно трудоемкий, а главное, длительный процесс, который чаще всего не дает желаемого результата. Решение этой проблемы на практике возможно путем создания автоматизированных систем сбора, фильтрации и анализа информации, так называемых «интеллектуальных посредников» между пользователем или корпоративной информационной системой и сетью Интернет.

Подобная система должна осуществлять сбор и селекцию информации из сети Интернет и создавать документальную базу данных, специфицированную предметной областью пользователя, то есть выполнять

функции интеграции информационных потоков.

Загрузка информации в БД должна сопровождаться ее классификацией и структуризацией. Для последующей информационно-аналитической работы пользователю должны предоставляться эффективные средства навигации, поиска и обобщения информации, которая сохраняется в соответствующей динамической документальной базы данных.

Анализ динамики тематических информационных потоков, которые генерируются в веб-пространстве, становится сегодня одним из наиболее информативных методов исследования актуальности тех или иных тематических направлений. Эта динамика обусловлена факторами, многие из которых не поддаются точному анализу. Однако общий характер временной зависимости количества тематических публикаций в сети Интернет все же допускает построение математических моделей.

В поведении информационных потоков наблюдаются две характерные особенности: во-первых, отчетливая тенденция к постоянному росту их объемов, а во-вторых, усложнение динамической структуры. Наблюдения временных зависимостей числа сообщений в сетевых информационных потоках убедительно свидетельствуют о том, что механизмы их генерации (генерирования) и распространения, очевидно, связаны со сложными нелинейными процессами общей сетевой динамики.

Анализируя задачи, связанные с моделированием информационных потоков и их влиянием на социальную среду, следует выделить несколько аспектов:

- структуру информационных потоков и их динамику;
- влияние информационных потоков на социальную среду;

- проблемы утечки или распространения информации в социальных группах;
- моделирование динамики потоков информации и ее обработки;
- особенности поведения систем в случае получения неполной, неточной или искаженной информации.

Попытки формализовать влияние информационных потоков на социальную среду, смоделировать поведение такой среды и отдельных ее элементов представлены в ряде работ [Додонов, 2009], [Ландэ, 2012], [Плотинский, 2006]. Развитие и верификация предложенных там моделей – это вопрос дальнейших исследований.

Возможностям моделирования информационных потоков, их структуре и влиянию на саму среду посвящены работы [Брайчевский, 2005], [Ландэ, 2006]. Особенно актуальным этот класс моделей может оказаться для анализа процессов взаимодействия людей через Интернет, который практически снимает ограничения на скорость передачи информации. Для изучения проблем распространения и утечки информации были предложены модели динамической перколяции, самоорганизующейся критичности.

При моделировании информационных процессов сохраняются все основные этапы моделирования:

- определение класса объектов, которые изучаются, построение модели, которая учитывает динамику развития социально правовых процессов;
- получение результатов моделирования и сравнение их с результатами наблюдений;
- установление соответствия модели и общественно-политической практики;
- анализ построенной модели и ее усовершенствование.

В настоящее время исследования по проблемам анализа информационных потоков большого объема в компьютерных сетях носят зачастую узко специализированный характер [Kowalczyk, 2014]. В то же время, опыт создания и внедрения ряда корпоративных информационных и информационно-аналитических систем, в частности, системы контент-мониторинга веб-ресурсов InfoStream [Григорьев, 2005], свидетельствует о необходимости создания и внедрения документальных информационных хранилищ для обеспечения научных исследований, получения различных аналитических сведений, навигации в документальных информационных потоках больших объемов.

Можно выделить две основные особенности объектов мониторинга. Первая из них – динамичность. Все объекты, исследование или обследование которых осуществляется с применением мониторинга, находятся в постоянном изменении, развитии. Вторая особенность – это наличие или возможность опасности, возникающей в процессе функционирования объекта мониторинга. Задачей мониторинга является предупреждение о том или ином риске, опасности, в широком смысле этого слова, для эффективного функционирования объекта. Причем не просто констатация факта появления изменений, которые представляют опасность, а именно предупреждение о ней – до того, как ситуация может стать необратимой. Тем самым создается возможность предотвратить или минимизировать возможное деструктивное развитие событий.

Динамичность объекта, возможность возникновения опасности в процессе его функционирования и размеры опасности определяют необходимость и целесообразность использования мониторинга для исследования, а также выбор той или иной конкретной системы мониторинга. Кроме этого необходимо отметить и еще одну особенность – возможность построения прогноза развития той или иной системы в условиях отсутствия флуктуационных

отклонений или форс-мажорных обстоятельств, что придает мониторингу особую ценность и значимость с точки зрения потенциального пользователя.

Попытки моделирования информационных потоков осуществлялись уже давно, но они тормозились вычислительными трудностями, особенно в случае необходимости описания динамики систем с обратными связями. Сегодня существует достаточное количество возможностей для эффективной компьютерной обработки текстовых данных, что позволяет, с одной стороны, готовить наборы входных параметров на основании анализа результатов статистических исследований, а с другой – решать формализованные задачи с достаточной степенью точности и в допустимое время.

Все это дает основание полагать, что в ближайшее время математическое моделирование станет основным инструментальным средством анализа и управления информационными потоками.

Одним из применений концепции эмерджентности к моделированию сегодня является многоагентное моделирование. Многоагентные модели широко применяются для анализа децентрализованных систем, закономерности динамики функционирования которых не изучены в достаточной мере. Эти модели используются в целях изучения общего поведения сложных систем, выявления правил их функционирования с учетом предположений об индивидуальном поведении ее отдельных компонентов.

В сложных системах (а современные ИС являются таковыми) среди многих других характеристик наиболее четко проявляется целостность, т.е. наличие таких свойств, которые не свойственны ни одному элементу (в данном случае, документу), что составляет систему, взятую отдельно. Это свойство, которое называют «эмерджентностью», является результатом возникновения между элементами системы особых синергетических связей. Под термином «эмерджентность», введенному Ф. Льюисом, понимается

то, что в системах целое является зачастую больше, чем сумма частей [7], то есть на каждом уровне сложности возникают новые, часто непредсказуемые качества, которые не свойственны составным частям.

Эмерджентность информационной системы не позволяет ограничиться изучением ее элементов и связей между ними, а допускает целостный анализ всей системы. К концу XX-го века при анализе сложных, в том числе и социальных, систем в основном использовался редукционистский подход, который объяснял множество свойств сложных систем свойствами их элементов – «атомов» или «молекул». В результате развития системного анализа, появления науки о сложности, технологического прорыва в вычислительных возможностях ситуация резко изменилась.

1.2.2. Формальное определение информационных потоков

Для поиска решения проблемы мониторинга, адаптивного агрегирования и обобщения документальных информационных потоков (ДИП) из глобальных компьютерных сетей необходимо формально с позиций системного анализа определить само понятие информационных потоков и исследовать их свойства. Результаты этого исследования должны составить теоретическую базу для разработки систем мониторинга, адаптивного агрегирования и обобщения информационных потоков, построения и ведения информационных ресурсов сверхбольших объемов и разнообразной тематической направленности.

Современный уровень развития информационного пространства обусловливает интерес к подходам, основанным на понимании информации как свойства упорядоченности некоторой системы и, соответственно, к статистическим методам ее обработки. Для организации эффективной коммуникации в сетях сегодня приходится постоянно возвращаться к истокам теории информации, понятиям энтропии, теории Шеннона, уравнениям Больцмана и др., широким

перспективам применения мощного аппарата математики и физики в решении теоретико-информационных задач.

Для формального описания информационных потоков введем некоторые общие для всего последующего изложения предположения. Дадим определение информационного потока [Додонов, 2009], которое корреспондируется с классическим определением в теории информации.

Рассмотрим отрезок действительной оси времени (a,τ), где $\tau > a$. Допустим, что на этом отрезке времени согласно некоторым закономерностям в сети публикуется некоторое количество документов – k. Пусть документы публикуются в моменты $\tau_1,\ \tau_2\ ,...,\ \tau_k\ \ (a \le \tau_1 \le \tau_2 \le ... \le \tau_k \le \tau)$.

Информационным потоком будем называть процесс $N_\alpha(\tau)$, реализация которого характеризуется количеством точек (документов), которые появились в интервале (a,τ), как функцию правого конца отрезка τ. Согласно этому определению реализация информационного потока является неубывающей ступенчатой целочисленной функцией $N_\alpha(\tau)$.

Приведенное определение на локальных временных областях соответствует действительности, но не учитывает такой эффект, как старение информации, которое противоречит «накопительной» способности информационного потока $N_\alpha(\tau)$ на больших промежутках времени.

Определенный таким образом информационный поток учитывает только количество информационных документов, независимо от их содержания. В общем случае определение содержания, тематики отдельных документов является достаточно субъективным процессом. Для строгого моделирования тематических информационных потоков используют модели, которые

различают документы по отдельным словам или словосочетаниям (обычно их называют термами, от англ. Terms).

Задачи мониторинга информационных потоков большого объема в глобальных компьютерных сетях, их адаптивной агрегации и обобщения усложняются отсутствием типовых методик и решений, неполнотой существующих технологических подходов. В настоящее время исследования по проблемам анализа информационных потоков большого объема в глобальных компьютерных сетях носят зачастую узко специализированный характер.

В то же время, опыт создания и внедрения корпоративных информационных и информационно-аналитических систем свидетельствует о необходимости создания и внедрения документальных информационных хранилищ для обеспечения научных исследований, получения различных аналитических сведений, навигации в документальных информационных потоках больших объемов.

При моделировании этих процессов используются методы нелинейной динамики и теории клеточных автоматов. При моделировании информационных потоков изучаются структурные связи между массивами документов, которые входят в них. При этом все чаще применяется фрактальный анализ, подход, основанный на свойствах сохранения внутренней структуры массивов документов при изменениях их размеров или масштабов рассмотрения.

Предусматривается, что новостные сообщения обладают свойством старения, т.е. теряют свою актуальность со временем. Все информационное пространство можно с достаточной мерой условности разделить на две составляющие – стабильную (статическую) и динамическую, которые имеют очень разные характеристики своего развития. В частности, процесс старения информации в известной модели Бартона-Кеблера [Bruton, 1960] описывается уравнением, которое состоит из двух компонент:

$$m(t) = 1 - ae^{-T} - b\mathrm{e}^{-2T},$$

где $m(t)$ – часть полезной информации в общем потоке через время T, первое вычитаемое соответствует стабильным ресурсам, а второе – динамическим, новостным.

Это уравнение также в полной мере соответствует объемам массивов информации, формирующимся в информационном пространстве по определенным тематикам, которые время от времени возникают и исчезают. Стабильная составляющая информационного пространства содержит информацию «долгосрочного» плана, в то время, как динамическая составляющая содержит ресурсы, которые постоянно обновляются.

Некоторая часть последней составляющей впоследствии вливается в стабильную, однако большая часть «исчезает» из информационного пространства или попадает в сегмент так называемой его «скрытой» части, не доступной пользователям с помощью обычных информационно-поисковых систем.

Теория информации, которая раньше находила свое основное применение в области передачи данных, становится полезной и для анализа текстовых массивов, которые динамично порождаются в сетях.

При исследовании информационных потоков как сложных систем широко используются экстремальные методы моделирования, которые применялись в естественных науках, в частности, на стыке экологии и биологии [Фурсова, 2003], [Webb, 1995], где они с используются при изучении популяционной динамики. Исследования, проводимые в этой области, почти без изменений могут применяться для изучения социальных процессов и информационных потоков.

Согласно экстремальным подходам к моделированию, реализуются только те состояния систем, которые отвечают экстремумам некоторой целевой функции (описываемой уравнениями) при определенных граничных условиях. Тончайшим

вопросом при этом являются принципы составления уравнений, которые в случае исследования информационных операций (как впрочем и в других областях) основываются на опыте экспертов, аналогиях, неполных эмпирических закономерностях.

При моделировании информационных потоков могут применяться подходы, основанные на логистических уравнениях роста популяций [Приц, 1974], [Lurie, 1983], [Левич, 1980], [Свирежев, 1991], [Ханин, 1982], [Quinn, 2014], получаемых в результате решения оптимизационных задач, принципов стационарного состояния открытых систем, максимального разнообразия популяции, максимальной обобщенной энтропии, максимума параметра мальтузианства и многих других.

1.2.3. Тематические информационные потоки

Под тематическим информационным потоком будем понимать последовательность сообщений, которые соответствуют определенной тематике. Таким образом, информационные системы в нашем понимании также являются тематическими информационными потоками, но в отличие от сообщений, которые проходят один за другим в простых информационных потоках, информационные системы – это сетевые структуры, охватывающие многочисленные информационные связи. В узком смысле под тематическим информационным потоком понимается количество документов, которые в некотором смысле соответствуют заданной теме.

Рассмотрим общую картину динамики тематических информационных потоков, ограничившись механизмами, типичными для динамического сегмента веб-пространства.

Многочисленные факты свидетельствуют о том, что на самом деле динамика тематических информационных потоков определяется комплексом внутренних нелинейных механизмов, которые лишь частично коррелируют с объективным окружением.

Очевидно, что эта динамика в принципе не может быть объяснена некоторым одним фактором, который полностью отвечает за все разнообразие эффектов, которые наблюдаются. Именно это обстоятельство и объясняет большую актуальность проблемы моделирования динамики тематических информационных потоков.

Информационный поток, который измеряется количеством сообщений, является величиной относительно стабильной. Изменяются во времени лишь объемы массивов сообщений, которые соответствуют той или иной тематике, той или иной информационной системе. Иными словами, рост количества публикаций по одной теме сопровождается уменьшением публикаций на другие темы [Ландэ, 2007], следовательно, для каждого промежутка времени T имеем:

$$\int_0^T \sum_{i=1}^{M} n_i(t)dt = NT ,$$

где $y_i(t)$ – количество публикаций в единицу времени по теме i, M – общее количество всех возможных тем. Обычно предполагается, что часть $n_i(t)$ всегда равна нулю. То есть, для локальных временных промежутков можно наблюдать так называемый «тематический баланс».

Основной интерес в такой формулировке представляет изучение динамики отдельного тематического потока, который описывается плотностью $n_i(t)$.

Теоретически можно допустить, что множество публикаций, которые ассоциируются с определенным набором тематик, пересекается, то есть существуют публикации, которые могут быть отнесены одновременно к нескольким различным тематикам. В реальности такая политематичность действительно наблюдается, она является фактором, который

необходимо учитывать, но в первом приближении считать, что его вклад не искажает общей картины.

Каждая тематика также имеет ряд характерных свойств, которые допускают некоторую классификацию, например, на основе особенностей его образования и воспроизведения во времени:

- публикации на «разовую» тему, временная зависимость количества которых резко растет, выходит на насыщение, а затем убывает и дальше асимптотически стремится к нулю;
- публикации по темам, которые периодически появляются в общем информационном потоке, а затем через некоторое время практически исчезают из него;
- публикации по темам, временная зависимость количества которых колеблется вокруг некоторого значения и никогда не исчезает полностью.

Таким образом, сообщения могут подразделяться на аналогичные категории, причем каждая из них имеет свою специфику развития во времени.

Еще сложнее выглядит синхронное изменение количества сообщений из нескольких тематических информационных потоков. Их поведение четко напоминает процессы взаимодействия популяций в биоценозе. Так, например, в ряде случаев увеличение количества публикаций по одной теме сопровождается сокращением количества публикаций на другие темы. Общая динамика в этом случае может описываться системой уравнений, каждое из которых относится к отдельному монотематическому потоку.

Подчеркнем, что общие политематические потоки являются стационарными по количеству публикаций, динамика же в основном определяется «конкурентной борьбой» отдельных тематик.

В то же время, в практическом плане часто

оказывается полностью удовлетворительным упрощенное понимание информационного потока как некоторой зависящей от времени величины $n(t)$, которая описывается уравнением:

$$\frac{dn(t)}{dt} = F\left(n(t), t\right).$$

В литературе описано много разновидностей систем «конкурентной борьбы» для различных модификаций модели в зависимости от целого ряда предположений относительно реальных условий протекания процессов. В простейшем виде такие уравнения могут иметь следующий вид:

$$\frac{dm_i(t)}{dt} = p_i \cdot m_i(t) - \sum_{j=1}^{N_m} r_{ij} \cdot m_i(t) \cdot m_j(t),$$

где N_m – количество тематик.

Приведенная система уравнений описывает перераспределение публикаций между тематиками, которые образуют фиксированный набор. Но в реальной жизни тематики (сюжеты) появляются и со временем исчезают, поэтому необходимо ввести в эти уравнения соответствующие коррективы. Это можно сделать по-разному, например, определив коэффициенты p_i и r_{ij} зависящими от времени так, чтобы каждый сюжет имел собственный максимум активности на определенном промежутке времени.

1.2.4. Синергетический подход к изучению информационных потоков

В конце XX-го века появилась новая междисциплинарная область науки, названная ее основателем Г. Хакеном «синергетикой» [Haken, 1977], [Haken, 1964]. Г. Хакен определяет синергетику как общую теорию коллективных пространственных, временных или функциональных макроструктур. Успехи современной синергетики в значительной степени

связаны с именами лауреата Нобелевской премии И. Пригожина, Г. Хакена, Г. Николиса, которым удалось свести множество идей, догадок, предположений о нелинейном характере причинности в единую методологическую концепцию-парадигму.

В основе эффективного моделирования информационных процессов в настоящее время применяются методы, которые предусматривают синергетические подходы. Известно, что синергетика (от греческого συν – "совместно" и εργος – "действующий") – это междисциплинарное направление научных исследований, задачей которого является изучение природных явлений и процессов на основе принципов самоорганизации систем.

В основе такого подхода лежит положение о том, что все участники правоотношения склоняются к переходу в определенные *равновесные точки (элементы самоорганизации)*, при переходе из одной точки в другую возможно несколько *альтернативных стратегий (точки бифуркации* в синергетике).

Физической системой, которая проложила путь синергетике, был лазер, потому что в физической теории этой открытой системы все концепции могут быть конкретно сформулированы и проверены. Затем, путем обобщения, была построена понятийная база синергетики [Вайдлих, 2005].

Лазер состоит из полости с параллельными зеркалами на двух противоположных сторонах и примерно тысячи активных атомов в этой полости. Эти атомы могут быть переведены из основного состояния в возбужденное каким-то внешним «накачивающим» источником. Обычно атомы, вследствие этого, совершают переход в основное состояние путем спонтанной эмиссии фотонов в случайных направлениях. Но когда количество этих возбужденных атомных состояний переходит определенный предел, происходит процесс, ведущий к динамическому фазовому переходу. В итоге получается лавина или поток фотонов (лазерный луч) макроскопического масштаба.

Синергетический подход близок к системному подходу, который лежит в основе большинства частных методов познания и является одним из способов обобщения фактов окружающей действительности.

Неупорядоченное, непрогнозируемое, случайное поведение системы связывают с недетерминированным хаосом, при котором невозможно вывести закономерности определения будущего состояния, зная ее предыдущее состояние. Сегодня все большее внимание ученых уделяется детерминированному хаосу, который порождается не случайным поведением большого количества элементов системы, а внутренней сущностью нелинейных процессов.

Поведение информационных систем в полной мере соответствует определению детерминированного хаоса. Для сложных систем, которыми, безусловно, являются такие системы, уравнения, описывающие их поведение, зачастую оказываются настолько сложными, что не могут решаться с помощью аналитических методов. Поэтому их исследование обычно проводится средствами компьютерного моделирования.

При решении нелинейных задач состояние системы и степень ее организованности изображают с помощью так называемого фазового пространства, координатами в котором являются параметры, которые характеризуют систему. Например, для описания систем в механике как координаты фазового пространства используются положение отдельных точек и их скоростей. В этом случае детерминированный хаос отображается непрерывной траекторией, которая временами может постепенно заполнять все фазовое пространство (любая малая окрестность точки в фазовом пространстве будет пересекать огромное количество фазовых траекторий). Это свойство детерминированного хаоса приводит к понятию фракталов, фрактальной размерности, например, хаусдорфова размерность траектории, которая плотно покрывает плоскость, не может быть целым числом. Дробная размерность – это один из основных признаков фракталов.

Основным предметом исследований для синергетики выступают процессы самоорганизации в сложных, открытых, неравновесных объектах-системах. Ее в первую очередь интересуют два типа трансформаций, через которые проходят сложные системы, включая социальные и информационные:

а) переходы от хаоса к порядку, то есть процессы возникновения новых форм, динамика самоорганизации в новообразуюшихся системах;

б) переходы от порядка к хаосу, то есть деструктивные процессы распада систем.

В настоящее время получили развитие такие направления, как теории хаоса, сложных сетей, нелинейных самоорганизующихся систем. Оказалось, что многие свойства сложных систем не могут быть выведены из заранее определенного набора динамических уравнений.

В то же время, очевидно, что невозможно разработать и применять на практике некоторую универсальную методику моделирования информационных систем. В основном это связано со слабой формализацией многих понятий и факторов, в первую очередь субъективных. В каждом отдельном случае приходится доверять информированности и интуиции аналитиков, профессионально занимающихся вопросами анализа информационных процессов.

С объективными факторами дело заключается иначе. Они полностью подвергаются анализу на статистическом уровне и допускают количественные оценки, которые могут использоваться для построения обоснованных прогнозов. Современные методы прикладной статистики, анализа временных рядов включают большой арсенал методов, которые детально проработаны и апробированы. Однако статистика позволяет описывать лишь формальные аспекты изучаемых явлений, оставляя за бортом аспекты содержательные. Поэтому существует необходимость расширения набора инструментальных средств,

используемых при анализе и моделировании информационных потоков.

Одним из наиболее перспективных направлений в этом плане является математическое моделирование. Сегодня математическое моделирование широко применяется во многих областях науки и техники, в то же время, моделирование информационных систем остается открытой проблемой.

В области информационных систем перспективным является моделирование, обусловленное некоторыми реалистическими правилами поведения отдельных элементов (документами, тематиками), которые уточняются некоторыми параметрами, которые изменяются при моделировании. В этом случае большую ценность имеет также и обратная задача – по реальному поведению некоторой зависимости оценить величину параметров модели.

Знание общего поведения устойчивых решений позволяет прогнозировать развитие ИС даже в том случае, когда не существует точного представления о конкретных механизмах, которые определяют их динамику, причем такого рода прогнозы могут оказаться точнее, чем полученные традиционными экспертными методами. Если же решения оказываются неустойчивыми, то из этого также может быть получена важная информация о системе, позволяющая в отдельных случаях прогнозировать, в какую сторону может быть направлена динамика отдельных ИС.

Под влиянием внешней среды информационные системы могут переходить к непредсказуемому поведению – хаосу. Неурегулированное, непрогнозируемое, случайное поведение системы связывают с недетерминированным хаосом, при котором невозможно вывести закономерности определения будущего состояния системы, зная ее предыдущее состояние. Сегодня все большее внимание ученые уделяют детерминированному хаосу, который порождается не случайным поведением большого

количества элементов системы, а внутренней сутью нелинейных процессов.

Ключевыми понятиями синергетики являются «бифуркации» и «аттракторы» [Капица, 1977]. Под точкой бифуркации обычно понимают состояние системы, после которого допустимо некоторое множество вариантов ее развития. Та траектория, или то множество траекторий, по которым возможно развитие системы после точки бифуркации, и которые отличаются от других относительной устойчивостью, называются аттракторами. То есть аттрактор якобы притягивает к себе множество траекторий, возможных после точки бифуркации. Свойства точек бифуркации и аттракторов изучаются в теории сложных систем, где устанавливаются закономерности развития таких систем, переходы от хаоса к порядку и наоборот.

Действительно, бифуркация может вызвать хаос. Опишем абстрактный, но достаточно убедительный пример, каскад бифуркаций М. Фейгенбаума (M. Feigenbaum), один из типичных сценариев перехода от простого периодического режима к сложному апериодическому при бесконечном удвоении периода [Feigenbaum, 1978].

Последовательность Фейгенбаума имеет самоподобную, фрактальную структуру – увеличение какой-либо области демонстрирует подобие выделенного участка со всей структурой.

Фейгенбаум в основном анализировал логистическое уравнение $X_{n+1} = CX_n - CX_n^2$, где C – внешний параметр, откуда вывел, что при некоторых ограничениях во всех подобных уравнениях происходит переход от равновесного состояния к хаосу.

Логистическое уравнение, которое, как известно, имеет два стойких решения, обычно трактуется как условие популяционной динамики и допускает следующую трактовку: предусматривается, что изолированно проживает популяция лиц численностью

X_n, через год появляется потомство численностью X_{n+1}. Рост популяции описывается первым членом правой части уравнения (CX_n), где коэффициент C определяет скорость роста и является определяющим параметром. Уменьшение числа животных (за счет перенаселенности, недостатка еды и т.п.) определяется нелинейным членом (CX_n^2). Результаты расчетов (рис. 2) показывают, что:

- при $C<1$ популяция при росте n вымирает;
- в области $1<C<3$ численность популяции приближается к постоянному значению $X_0=1-1/C$, которое определяет область стационарных, фиксированных решений. При значении $C=3$ точка бифуркации становится отталкивающей фиксированной точкой;
- в диапазоне $3<C<3.57$ начинают появляться бифуркации – разветвления каждой кривой на две (действительно, логистическое уравнение имеет два стойких класса решений). Численность популяции колеблется между двух значений, которые лежат в этих областях. Сначала популяция резко растет, на следующий год возникает перенаселенность и через год численность опять уменьшается;
- при $C>3.57$ происходит перекрытие разных областей и поведение системы становится на вид хаотическим.

Таким образом, заключительным состоянием физических систем, которые эволюционируют, является состояние динамического хаоса.

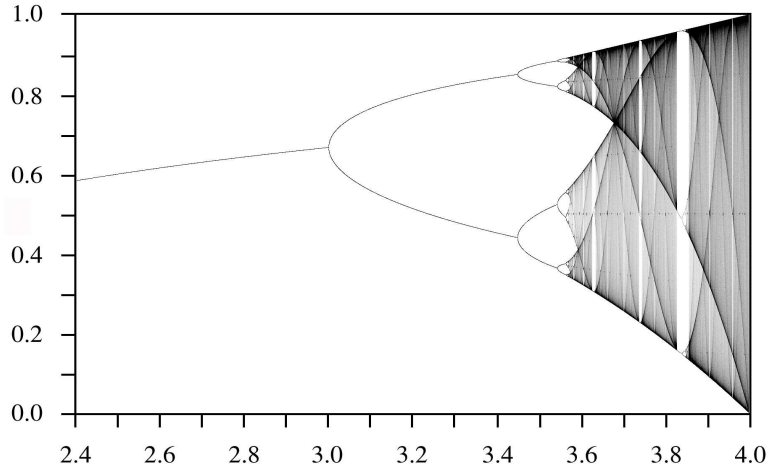


Рис. 2 – Каскад бифуркаций (последовательность Фегенбаума): ось абсцисс – значения параметра C, ось ординат – значения X_n

С помощью хаоса не только нельзя построить или уточнить прогноз, но и, соответственно, проверить его. Однако это не должно говорить о неверности теории хаоса, подтвержденной как в математических расчетах, так и в жизни. Сейчас еще не существует математически точного применения теории хаоса для исследований информационных потоков, но в то же время, эта теория уже сегодня способна предусмотреть переходы моделей систем, представленных в аналитическом виде, в хаотическое состояние. Очевидно, это действительно одно из самых перспективных направлений прикладных исследований информационных процессов.

В синергетике строго доказывается, что никакими внешними воздействиями невозможно «навязать» системе желаемую кому-нибудь поведение – можно лишь выбрать наилучшую соответствующую траекторию из потенциально заложенных [Потеев, 1999]. Поэтому при планировании и моделировании ИС одной из основных задач является нахождение точек

бифуркации информационных процессов и формирование флуктуаций, приводящих к выбору необходимой траектории эволюции (к аттрактору).

Как известно, в математике катастрофами называются скачкообразные изменения, возникающие в виде ответа системы на плавное изменение внешних условий. Информационные системы могут вызвать процессы, которые лучше всего описываются в рамках теории катастроф: «около точек бифуркации в системах можно наблюдать значительные флуктуации. Такие процессы будто колеблются перед выбором одного из нескольких путей эволюции. Небольшая флуктуация может служить началом эволюции в полностью новом направлении, которое резко изменит все поведение макроскопической системы» [Пригожин, 1986]. Это объясняет, почему так трудно бороться с катастрофой, когда ее признаки стали уже заметными: скорость ее приближения бесконечно растет по мере приближения к катастрофе [Арнольд, 1990].

Если рассматривать общество как сложную систему, то информационные системы можно рассматривать как методы воздействия на эту систему, как было показано выше с целью выбора определенных путей развития. Таким образом модели информационных систем являются частью общих социальных моделей.

1.3. Источники информации

В информационно-аналитической работе важное значение имеет возможность доступа к источникам информации. При этом главной проблемой является нахождение содержательных и надежных источников из всех общедоступных.

Если ранжировать количество источников, которые можно получить при применении трех приведенных выше подходов, то, вероятно, можно в очередной раз получить подтверждение общенаучной закономерности Брэдфорда, которая, в свою очередь, вытекает из закона Ципфа. Закономерность Брэдфорда

в первоначальном виде относилась к традиционным «бумажным» изданиям.

Исследуя различные типы источников информации в области науки, С. Брэдфорд [Bradford, 1934] распределил их по трем множествам, равными по количеству полезной информации документов: R_1, R_2, R_3. Здесь R_1 – это самые рейтинговые источники, которые непосредственно относятся к определенной тематике; R_2 – множество источников, которые корреспондируют с компьютерной тематикой; R_3 – источники, которые частично касаются данной темы.

При этом количество полезной информации во всех трех множествах является постоянной.

Если принять обозначение $|A|$ – количество элементов множества A, то пропорция Брэдфорда записывается следующим образом:

$$| R_1 | : | R_2 | : | R_3 | = C.$$

Для множеств документальных источников в сети Интернет справедливы соотношения:

$$|S_1| : |S_3| = |S_3| : |S_2| = C,$$

где S_1 – это множество источников, полученных по алгоритму 1 (выбраны источники с веб-пространства); S_2 – множество источников, которые корреспондируют с алгоритмом 2 (поиск в глобальных сетевых ИПС); S_3 – источники, которые соответствуют алгоритму 3 (применение специализированных систем поиска и баз данных); C – некоторая константа, которая соответствует информационным потребностям пользователей.

Когда необходимые для проведения аналитического исследования источники найдены, включаются механизмы превращения информации в знания, для чего применяются соответствующие информационные технологии. Конечным продуктом любой аналитической работы являются знания –

синтезированные выводы, рекомендации для принятия решений.

Информация в глобальной сетевой среде может быть получена из открытых источников, рекламы, фирменных, банковских, правительственных отчетов, баз данных, от экспертов путем анализа или специальной обработки данных, текстов.

Ниже приведен примерный список видов информационных источников, которые чаще всего используются в аналитической работе [Нежданов, 2009]:

1. Пресс-релизы компаний, официальные заявления от имени компаний о новых технологиях, новых направлениях, сделках, перспективах. Такие пресс-релизы создаются компаниями для собственной популяризации, привлечения внимания потенциальных клиентов, инвесторов, ищущих выгодные варианты вложения своих средств. Часто в таких заявлениях присутствует информация о намерениях, планируемых событиях. Пресс-релизы доступны на веб-сайтах компаний, в PR-службах, на общих и профильных специализированных площадках для размещения пресс-релизов.

2. Интервью сотрудников компаний, соответствующие материалы в СМИ. В интервью особый интерес представляют планы компаний. При этом со стороны службы конкурентной разведки допускается инициирование интервью кого-то из сотрудников объекта интереса.

3. Высказывания сотрудников компаний на форумах, в блогах. При этом могут выявляться планы компаний, кадровая политика, атмосфера в коллективе и т. п. Источники информации: 1) интернет-ресурсы (специализированные форумы, блоги сотрудников), блоги экспертов, группы в социальных сетях; 2) выставки, конференции, курсы повышения квалификации, профессиональные мероприятия.

4. Тендеры, закупки. Предметы закупок, оборудование, исполнители. Источники информации: 1) интернет-ресурсы (веб-сайты компаний, торговые площадки, профильные форумы); 2) партнеры исследуемой компании, те, кто участвовал в их тендерах, у клиентов и поставщиков.

5. Патенты, авторские свидетельства компании и ее сотрудников. Для задач сетевой аналитики зачастую интересно их содержание, направленность, списки соавторов.

6. Разработки компании: ведущиеся, финансируемые, разработки, которыми компания интересуется. Наблюдению подлежат попытки компании проводить исследования: закупка специфического оборудования, прием на работу специалистов, переговоры, посещения соответствующих организаций и т.д.

7. Активность компании на рынке слияний и поглощений (M&A). Информация о том, какие организации поглощаются, планируют поглотить или ведут переговоры о поглощении. Информацию можно получить, например, в Антимонопольном комитете (АМК) Украины, в Федеральной антимонопольной службе Российской Федерации (ФАС России), по новостным сообщениям на веб-ресурсах посвященных M&A.

8. Вакансии компании (открывающиеся, закрывающиеся), сообщения об активном поиске сотрудников, требования к вакансиям, условия. Источник информации: веб-сайт компании, сайты по поиску работы и веб-сайты агентств, с которыми компания сотрудничает.

9. Благодарности и награды компании и ее сотрудников.

10. Веб-ресурсы мероприятий (выставок, конференций и т.п.). Выяснение, в каких мероприятиях участвуют компании, их направленность, круг участников.

12. Веб-ресурсы организаций (союзы, ассоциации, академии и т. п.) – информация о том, в каких объединениях участвуют компании, как активно они участвуют, что получают от участия, на что рассчитывают, как используют.

Информация характеризуется качественными, количественными и ценностными показателями. К качественным характеристикам обычно относят: достоверность, объективность и однозначность информации. К количественным характеристикам – ее полноту (отсутствие невыясненных пробелов) и релевантность (степень соответствия существу поставленных вопросов и задач). Ценностными характеристиками являются стоимость и актуальность информации.

Многие аналитические службы не всегда могут отделить нелегитимную часть информации от легальной, а заказчик, как правило, интересуется конечными результатами, источники для него выступают лишь в качестве подтверждений, промежуточных данных. Вместе с тем, солидные заказчики сами заинтересованы в том, чтобы информация добывалась законными средствами, чтобы аналитический отчет был легален.

У аналитических служб в последние десятилетия появился и развился до невиданных ранее масштабов новый информационный источник – веб-пространство сети Интернет. Сегодня по оценкам экспертов Интернет по количеству информации находится на первом месте, опережая СМИ, отраслевые издания и новости, специальные обзоры, закрытые базы данных. При этом в открытых источниках и специализированных базах данных, доступных в Интернет, содержится большая часть информации, необходимой для аналитической деятельности, однако остается открытым вопрос ее нахождения и эффективного использования. Последние исследования информационного веб-пространства показали, что доступный через традиционные информационно-поисковые системы триллион веб-

страниц – это лишь «поверхностная видимая часть айсберга». Около 40% всей информации в Интернете доступно бесплатно. Навигацию по данному информационному пространству обеспечивают более миллиона поисковых систем и каталогов, но и они охватывают лишь малую часть информационных ресурсов. Скрытых и невидимых (deep, invisible) ресурсов сети Интернет значительно больше – это прежде всего динамически-генерируемые страницы, файлы разнообразных форматов, информация из многочисленных баз данных [Devine, 2013]. К «скрытому» веб можно отнести и такие сети, как BitTorrent, DirectConnect, EMule, Napster и др.

Сегодня при проведении аналитической деятельности основными источниками информации служат Интернет, пресса, а также открытые базы данных. Очень популярны среди аналитиков базы данных государственных и статистических органов, торгово-промышленных палат, органов приватизации и т.д. Большую пользу приносят и отдельные доступные базы данных других органов государственной власти. В последнее время все более популярны базы данных на основе архивов СМИ, базы данных на основе архивов масс-медиа, в том числе (и преимущественно) сетевых. В постсоветских странах, например, большой популярностью пользуется крупнейшая архивная база данных СМИ службы «Интегрум» (integrum.ru), ретроспективный фонд сервиса InfoStream (infostream.ua), содержащие несколько сотен миллионов документов. С помощью базы данных «Лабиринт» (labyrinth.ru), составленной на основе публикаций ведущих бизнес-изданий, можно получить обширную информацию о конкретных персонах, организациях и компаниях.

1.3.1. Веб-пространство как сложная сеть

Веб-пространство базируется на физической инфраструктуре сети Интернет и протоколе передачи данных HTTP. Количество веб-сайтов в сети Интернет –

919,533,715 (по состоянию на март 2014 года по данным службы Netcraft, рис. 3).

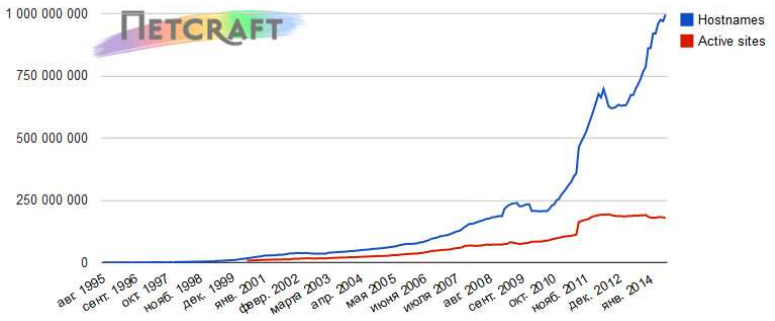


Рис. 3 – Динамика роста количества веб-серверов в логарифмической шкале (Netcraft, январь 2013 года)

Однонаправленные связи между отдельными веб-страницами реализуются в виде гиперссылок (рис. 4).

В начале существования веб-пространства на небольшом количестве веб-сайтов публиковалась информация отдельных авторов для относительно большого количества посетителей. Сегодня ситуация резко изменилась, произошел переход к вебу второго поколения. Сами посетители веб-сайтов активно участвуют в создании контента, что привело к резкому росту объемов информации и динамики веб.

Сегодня в веб уже существует свободно доступная для пользователей информационная база такого объема, который ранее трудно было представить. Более того, объемы этой базы превышают на порядки все то, что было доступно десятилетие назад. В августе 2005 года компания Yahoo! объявила о том, что проиндексировала около 20 млрд. документов. Достижение компании Google в 2004 году составляло менее 10 млрд. документов. Сегодня Google заиндексировала свыше триллиона веб-документов. По данным службы Netcraft Web Server Survey (news.netcraft.com), в настоящее время количество веб-серверов превышает 670 млн.

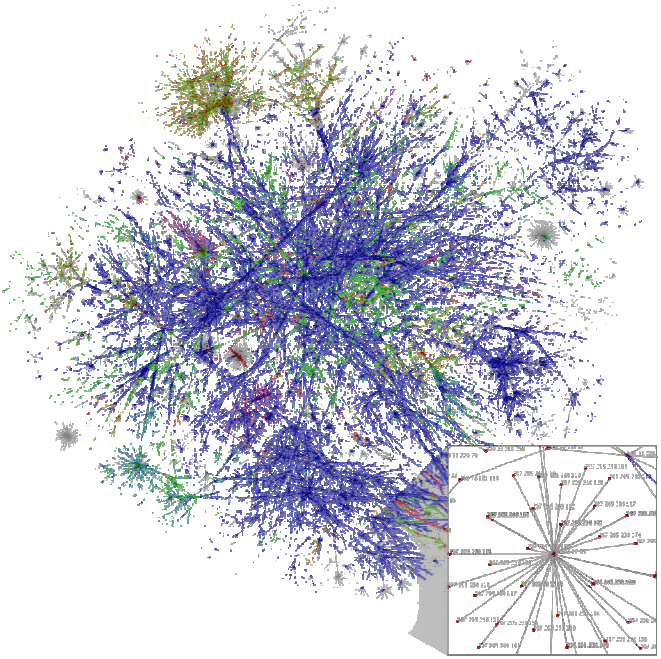


Рис. 4 – Карта связей Интернет-серверов как сложная сеть (по данным en.wikipedia.org)

В открытых источниках и специализированных базах данных, доступных в веб-пространстве, содержится большая часть информации, необходимой для проведения аналитических исследований, однако остаются открытыми вопросы ее нахождения и эффективного использования. При использовании веб-пространства как мощнейшего источника информации, как уже было отмечено ранее, самыми существенными являются проблемы объема, навигации, наличия информационного шума и динамического характера информации в Интернет.

Возможности доступа к интернет-ресурсам, привлекающим своей открытостью, объемами и содержательной многогранностью, на первый взгляд кажутся безграничными. Однако важные события в различных областях свидетельствуют об обратном.

Именно в кризисных ситуациях Интернет довольно часто подводит. Существует множество проблем – от перегруженности сетевой инфраструктуры – до вирусных атак, уязвимостей и отказов в обслуживании отдельных веб-серверов. Целый ряд проблем порожден также объемами, разнообразием представления и динамикой контентного сегмента информационного пространства.

Несмотря на такие качества, как открытость и доступность, существующую инфраструктуру веб-пространства нельзя признать надежной и достоверной.

Назовем еще несколько проблем, присущих веб-пространству:

- не решена задача доступа пользователей к разнородным веб-ресурсам из «одного окна» для получения обобщенного представления потоков информации по необходимой тематике;
- не обеспечена возможность своевременного «напоминания» и «проталкивания» профильной для пользователя информации, публикуемой на большом количестве веб-сайтов;
- достаточно большая вероятность отказа в обслуживании критически важных веб-ресурсов в самое неподходящее время.

Известно, что сегодня существуют технологии интеграции контента, которые позволяют частично решать названные проблемы, обеспечивая эффективный поиск и навигацию в веб-пространстве, мониторинг и агрегацию открытых веб-ресурсов.

Для профессионального поиска в веб-пространстве и мониторинга информации используются специализированное программное обеспечение, информационно-поисковые системы и сервисы. Приведем некоторые примеры программных продуктов:

Copernic Agent (www.copernic.com/en/products

/agent) – программа, позволяющая проводить метапоиск, используя, как заявлено на веб-сайте компании, 1000 поисковых систем, объединять результаты, устранять дубликаты, блокировать нерабочие ссылки, показывать наиболее релевантные результаты.

Avalanche (www.tora-centre.ru) – семейство программных средств для веб-мониторинга. Технология Avalanche базируется на трех основных решениях: концепции «умных папок» (Smart Folders), автономном интеллектуальном поисковом роботе и встроенной базе данных («персональной энциклопедии»).

Newprosoft Web Content Extractor (www.newprosoft.com) – программа сканирования и извлечения данных из веб-сайтов.

Portable Offline Browser от MetaProducts Corporation (www.portableofflinebrowser.com) – программа, позволяющая скачивать необходимые веб-сайты и мультимедийную информацию, в том числе Flash-анимацию, скрипты и активное содержимое страниц.

Neiron Search Tools (neiron.ru/toolbar) – программная надстройка, объединяющая результаты информационно-поисковых систем Google и Яndex, которая позволяет осуществлять конкурентный анализ, базирующийся на оценке эффективности сайтов и контекстной рекламы.

WebSite-Watcher (www.aignes.com) – программа, позволяющая проводить мониторинг веб-сайтов, форумов, локальных файлов, обеспечивающая фильтрацию информации, а также удобную визуализацию результатов мониторинга.

В качестве сервисных решений можно назвать:

WatchThatPage (watchthatpage.com) – бесплатный сервис, позволяющий автоматически собирать новую информацию с веб-ресурсов, поставленных на мониторинг.

Diphur Monitor EveryThing (www.diphur.com) – бесплатный сервис мониторинга любых веб-сайтов, уведомляющий об их обновлении и доставляющий пользователям обновления.

Newspaper Map (newspapermap.com) – сервис, объединяющий геолокацию и информационно-поисковую систему по медиа-ресурсам. При решении задач конкурентной разведки пользователь может выбрать интересующий его регион, язык, список онлайн версий газет и журналов, непосредственно выходить на документы. Сервис поддерживает русский язык, имеет удобный интерфейс.

InfoStream (www.infostream.ua) – сервис контент-мониторинга веб-ресурсов России и Украины, предоставляющий доступ в поисковом режиме к информации из 7000 источников, классификацию информации, экстрагирование понятий (персон, компаний, топонимов), формирование сюжетных цепочек, оценку тональности сообщений, анализ динамики публикаций по определенным объектам (рис. 5).

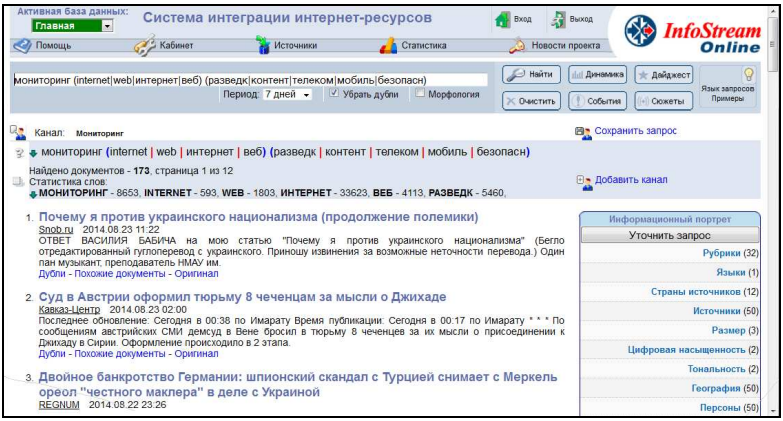


Рис. 5 – Интерфейс системы контент-мониторинга InfoStream

Agregator.pro (agregator.pro) – агрегатор информации с новостных и медийных порталов. Может

использоваться в конкурентной разведке для отслеживания интересующих объектов, получения частоты и контекста упоминания отслеживаемого объекта в СМИ, анализа динамики обращений по времени.

WebGround (webground.su) – агрегатор новостной информации из русскоязычного сегмента веб-пространства. Может использоваться в конкурентной разведке для отслеживания интересующих тематик, получения тематических сюжетов, ретроспективного анализа развития тематики во времени. Фрагмент интерфейса агрегатора новостей WebGround представлен на рис. 6.

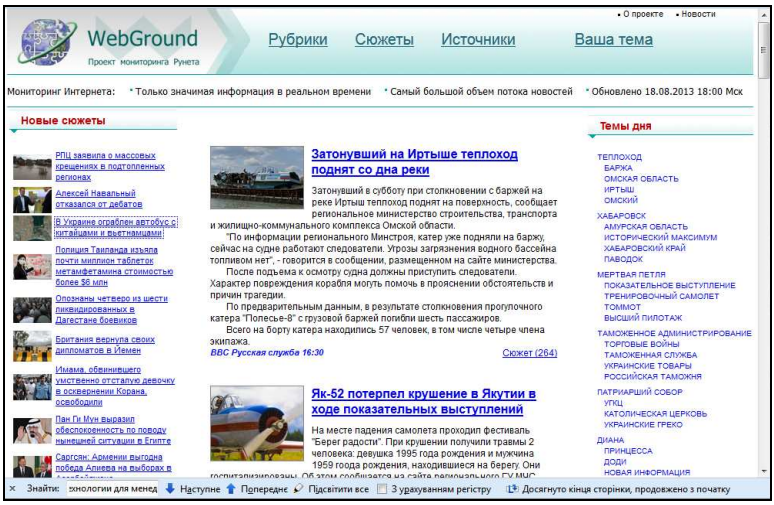


Рис. 6 – Фрагмент интерфейса системы WebGround

1.3.2. Глубинный веб

Последние исследования веб-пространства показали, что доступные через традиционные информационно-поисковые системы более триллиона веб-страниц – это лишь «поверхностная видимая часть айсберга». Важной проблемой является поиск информации в «скрытом» или «глубинном» веб-пространстве, где, как было замечено выше, содержится

несравнимо большее количество данных, потенциально интересных для конкурентной разведки, чем в открытой части Интернета

Это, прежде всего, динамические веб-страницы, информация из многочисленных баз данных, которые могут представлять большой интерес для аналитической работы. К разряду «скрытого» веб относятся и полнотекстовые информационные системы типа LexisNexis или Factiva.

К «скрытым» ресурсам сети Интернет можно отнести также пиринговые сети, такие как BitTorrent, EDonkey, EMule, Gnutella, Kazaa.

Как уже было отмечено ранее, необходимой (в том числе и для конкурентной разведки) информации в сети Интернет значительно больше, чем ее охватывают универсальные поисковые машины. Предполагается, что в отличие от «познаваемой» части сети Интернет, «скрытая» часть оказалась в сотни раз более объемной.

Бизнес-аналитик часто сталкивается с ситуацией, когда ему известно о существовании в веб-пространстве какого-то документа, но не может найти его с помощью традиционных поисковиков, какими сегодня можно считать такие системы, как Google, Yahoo!, Bing, Яндекс, Рамблер или Мета. Однако, вспомнив или найдя в закладках адрес (URL) этого документа, он без труда выходит на него. То есть в веб-пространстве этот документ есть, а найти его привычным способом нельзя. Пользователь столкнулся с невидимым (*invisible*) для поисковых систем ресурсом.

Что такое глубинный веб?

Совокупность источников в веб-пространстве, недоступных пользователям традиционных поисковых систем, образует так называемый «глубинный веб» – понятие, введенное Джиллом Иллсвортом (Jill Ellsworth) в 1994 г. Под глубинным веб (invisible web, deep web, hidden web) принято понимать ту часть веб-пространства, которая не индексируется роботами (web

crawlers) поисковых систем. Используя аналогию, информация, будучи недоступной для поиска, находится «в глубине» (англ. – deep). При этом не стоит путать глубинный веб с ресурсами, вовсе недоступными из сети Интернет – это темный веб (dark web), и речь о нем здесь идти не будет. Некоторые ресурсы, доступ к которым открыт лишь для зарегистрированных пользователей, также относятся к глубинному веб.

В 2000-м году американская компания BrightPlanet (www.brightplanet.com) опубликовала сенсационный доклад, в котором утверждается, что в веб-пространстве в сотни раз больше страниц, чем их удалось проиндексировать самыми популярными на то время поисковыми системами. Компания разработала программу LexiBot, которая позволяет сканировать некоторые динамические веб-станицы, формируемые из баз данных, и запустив ее получила неожиданные данные. Выяснилось, что в глубинном веб находится в 500 раз больше документов, чем доступно через поисковые системы. Конечно, эти цифры неточны. Кроме того, стало известно, что средняя страница глубинного веб на 27% компактней средней страницы из видимой части веб-пространства.

Сегодня ситуация изменилась, например, ведущие поисковые системы могут индексировать документы, представленные в форматах, содержащих текст. Конечно, это прежде всего pdf, rtf и doc. В 2006-м г. Google запатентовала способ поиска в глубинном веб: «Searching through content which is accessible through web-based forms» (рис. 7). По мнению разных исследователей к видимому веб относится лишь 20-30% веб-пространства.

Причины возникновения

В глубинном веб находятся веб-ресурсы, не связанные с остальными ресурсами гиперссылками – например, страницы, динамически создаваемые по запросам к базам данных, документы из баз данных, доступные пользователям через поисковые веб-формы (но не по гиперссылкам). Такие документы остаются

недоступными для робота, неспособного в режиме реального времени правильно заполнить поля формы значениями (формировать запросы к базам данных).

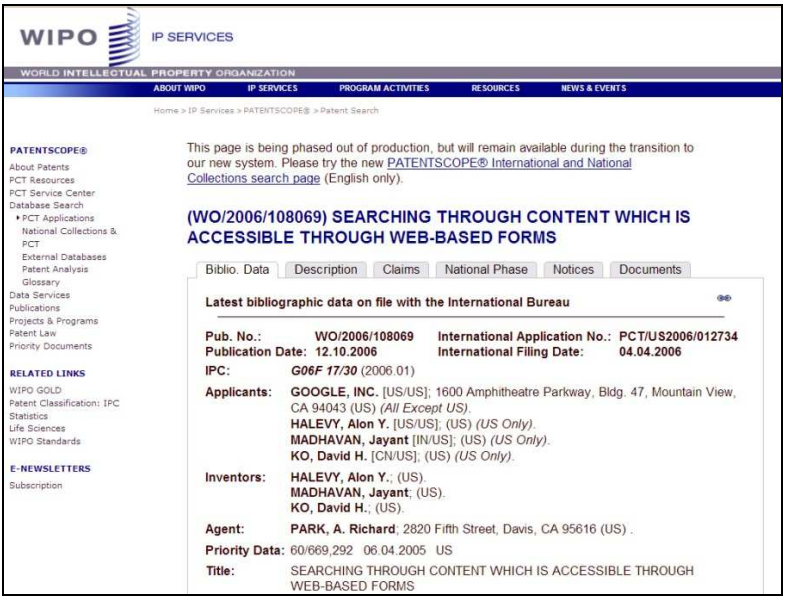


Рис. 7 – Фрагмент веб-ресурса WIPO с описанием патента Google на поиск в глубинном веб

Вот что говорится о глубинном веб в книге [Price, 2001]: «Большинство страниц невидимого Интернета могут быть проиндексированы технически, но не индексируются, потому что поисковые системы решили их не индексировать… Большинство «невидимых» сайтов имеют высококачественный контент. Просто эти ресурсы не могут быть найдены с помощью поисковых машин общего назначения…».

Некоторые сайты используют технологию баз данных, что действительно сложно для поисковой машины. Другие сайты, однако, используют сочетание файлов, которые содержат текст и мультимедиа, а поэтому часть из них может быть проиндексирована, а часть – нет.

Многие сайты глубинного веб могут быть проиндексированы поисковыми машинами, но это не делается потому, что поисковые машины считают это непрактичным – например, по причине стоимости или потому, что данные настолько короткоживущие, что индексировать их просто бессмысленно – например, прогноз погоды, точное время прибытия конкретного самолета, совершившего посадку в аэропорту и т.п.»

Основные ограничения, связанные с роботами поисковых машин можно объяснить следующими основными причинами: для публичных поисковых служб важнее обеспечить точность поиска, чем полноту, важнее обеспечить получение ответа на запрос в приемлемое время, чем точность. Отсюда – ограничения на глубину сканирования веб-ресурсов, попытки «фильтрации» контента по содержанию, отсеивание страниц, содержащих излишние выходные гиперссылки и т.п. При этом часто с водой выплескивается и ребенок. Общепризнано, что ценность ресурсов глубинного веб зачастую выше ценности ресурсов видимой части веб-пространства.

Можно упомянуть еще один источник пополнения глубинного веб – владельцы сознательно не хотят чтобы их веб-ресурсы находили с помощью поисковых систем. Чаще всего такие веб-ресурсы представляют нечто не совсем законное, хакерские форумы, архивы неавторизованного контента и т.п. Понятно, что многие из таких ресурсов очень интересны для изучения бизнес-аналитиками.

Многие компании сначала подключаются к сети Интернет, и лишь потом тратят большие средства на защиту. Владельцы веб-сайтов могут попытаться запретить индексацию тех или иных страниц своих ресурсов, прописав запрещающую команду в файле robots.txt, но поисковые системы могут ее проигнорировать. Поэтому такие ресурсы либо удаляют, либо удаляют гиперссылки, переводя ресурсы в категорию глубинного веба.

Виды ресурсов глубинного веб

Существует несколько типов ресурсов глубинного веб, например, как было отмечено выше, это могут быть быстро устаревающие веб-страницы. Кроме того, к глубинному веб относятся веб-ресурсы, представляющие собой мультимедийную информацию. Как известно, в данное время еще не существует удовлетворительных алгоритмов поиска нетекстовой информации. Динамически генерируемые по запросу страницы также часто попадают в глубинный веб. Зачастую без запроса таких страниц не существует, они генерируются при запросе к базам данных. Получается, что информация, вроде бы и присутствует в веб-пространстве, но возникает она лишь в момент обработки запроса, а универсального алгоритма заполнения их роботами поисковых форм не существует. И, наконец, если на веб-ресурс не ведут никакие ссылки, то роботы поисковых систем никаким образом не могут узнать об его существовании.

Основатель компании BrightPlanet Майкл Бергман (Michael K. Bergman) смог выделить 12 разновидностей глубинных веб-ресурсов, относящихся к классу онлайновых баз данных. В списке оказались как традиционные базы данных (патенты, медицина и финансы), так и публичные ресурсы – объявления о поиске работы, чаты, библиотеки, справочники. Бергман причислил к глубинным ресурсам и специализированные поисковые системы, которые обслуживают определенные отрасли или рынки, базы данных которых не включаются в глобальные каталоги традиционных поисковых служб.

К глубинному веб также относятся многочисленные системы интерактивного взаимодействия с пользователями – помощи, консультирования, обучения, требующие участия людей для формирования динамических ответов от серверов. К ним также можно отнести и закрытую (полностью или частично) информацию, доступную, пользователям Сети

только с определенных адресов, групп адресов, иногда городов или стран.

К «скрытой» части Сети многие причисляют и веб-страницы, зарегистрированные на бесплатных серверах, которые индексируются, в лучшем случае, лишь частично – поисковые системы во избежание рекламного спама не стремятся обходить их в полном объеме.

К глубинному веб также относится категория так называемых «серых» сайтов, функционирующих на основе динамических систем управления контентом (Dynamic Content Management Systems). В поисковых системах обычно ограничивается глубина индексирования таких сайтов во избежание возможного циклического просмотра одних и тех же страниц.

Примеры ресурсов глубинного веб

Как же найти веб-ресурсы, размещенные в глубинном веб? Если ресурсы требуют заполнения специальных форм, дополненных, например, капчами, то необходимо выйти на базу данных, предположительно содержащую необходимые документы. Найти базы данных – источники скрытого веб можно с помощью обычных поисковых систем, обобщив запрос и введя уточняющие слова, такие как «база данных», «банк данных», «database» и т.п.

Приведем общеизвестный пример: пользователю требуется статистика по катастрофам самолетов в Аргентине. Естественный запрос к традиционной поисковой системе выдает огромный список газетных заголовков. На запрос «aviation database», можно сразу выйти на базу данных NTSB Aviation Accident Database (www.ntsb.gov/ntsb/query.asp).

Для поиска в глубинном веб, а именно в том его сегменте, который составляют базы данных, сегодня уже существуют некоторые специализированные ресурсы. Лидером среди навигаторов в глубинном веб является сайт CompletePlanet (www.completeplanet.com)

компании BrightPlanet. Этот сайт является крупнейшим каталогом, насчитывающим свыше 100 тысяч ссылок. Компания BrightPlanet также создала персональную утилиту для поиска в онлайновых базах данных LexiBot, которая может обеспечивать поиск в нескольких тысячах поисковых систем «глубинного» веб. Метапоисковый пакет DeepQueryManager (DQM) этой же компании обеспечивает поиск более чем по 70 тысячам «скрытым» веб-ресурсам.

Исследование, проведенное еще в 2006 г. [Не, 2007] показало, что глубинный веб охватывает более 300 тыс. сайтов, связанных с более 450 тыс. базами данных, не охватываемых традиционными поисковыми системами. К наиболее интересным для бизнес-аналитиков ресурсам глубинного веб относятся: базы данных юридических и физических лиц; отраслевые базы данных; репутационные базы данных (черные и белые списки); криминологические базы данных; базы данных товаров и услуг; каталоги продукции и т.п. К всемирно известным бизнес-ресурсам, размещенным в глубинном веб относятся: amazon.com, ebay.com, realtor.com, cars.com, imdb.com.

Приведем еще несколько примеров баз данных и каталогов глубинного веб:

FindLaw (www.findlaw.com) – один из наиболее популярных в мире юридических веб-сайтов – огромный каталог правовых ресурсов, содержащий аннотированный список свободно доступных баз данных нормативно-правовых документов, для которых данный ресурс является «точкой входа» (рис. 8).

About.com (www.about.com) – портал, охватывающий тысячи, снабженных комментариями, ссылок на веб-ресурсы, в том числе и на ресурсы глубинного веб (имеется ссылка «Invisible Web»). На портале предоставляется возможность поиска в каталоге. Ресурс также включает несколько статей по проблематике глубинного веб: «What is the Invisible Web?», «Finding the Invisible Web», «Top Places to Search the Invisible Web» и др.

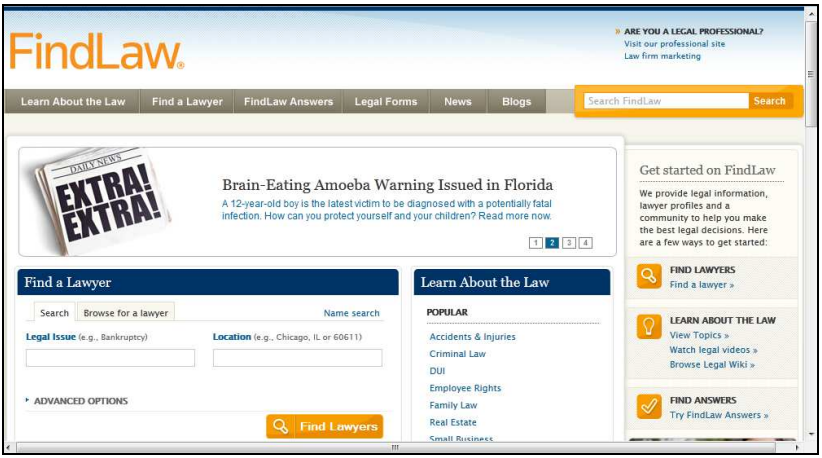


Рис. 8 – Фрагмент веб-сайта сервиса FindLaw

Politicalinformation.com (www.politicalinformation.com) – сервис, обеспечивающий оперативный поиск в 5000 отобранных веб-сайтах политической направленности, предоставление новостей из нескольких десятков авторитетных источников.

Infomine (infomine.ucr.edu) – сервис обеспечивает добычу информации из баз данных, электронных журналов (блогов), электронных досок объявлений, электронных книг, списков рассылок, электронных каталогов и т.п., преимущественно познавательно-образовательного характера. Обеспечивает как общий поиск, так и поиск по тематическим категориям.

Особенность большинства «скрытых» ресурсов заключается в их узкой специализации. Для поиска в них используются те же механизмы, что и для «поверхностного» веб, однако, в большинстве случаев, роботы поисковых систем для глубинного веб включают уникальные для каждого такого ресурса модули доступа к данным.

Традиционная поисковая система чаще всего может выдать адрес базы данных, но не скажет, какие документы конкретно содержатся в ней. Типичный пример – информационно-поисковые системы по

украинскому (zakon.rada.gov.ua) или российскому законодательству (www.kodeks.ru). Тысячи документов из баз данных становятся доступны только после входа в систему, а роботы стандартных поисковых систем не в состоянии заиндексировать контент баз данных.

Парадоксально, но в качестве одного из ресурсов глубинного веб можно рассматривать и архив ресурсов открытого веб-пространства. Такой архив – Internet Archive с 1996 года создает компания Alexa (www.archive.org). Сегодня объем базы данных Alexa превышает 350 млрд. веб-страниц (рис. 9).

Технология хранилища Alexa включает ряд современных средств управления гигантским документальным хранилищем. Например, с помощью технологии Alexa выполняется кластеризация веб-ресурсов, т.е. формирование коллекций документов, близких по тематикам. Особый интерес у пользователей сервиса Alexa вызывает «Машина времени» (Wayback Machine), открывающая доступ к временным срезам веб-пространства.

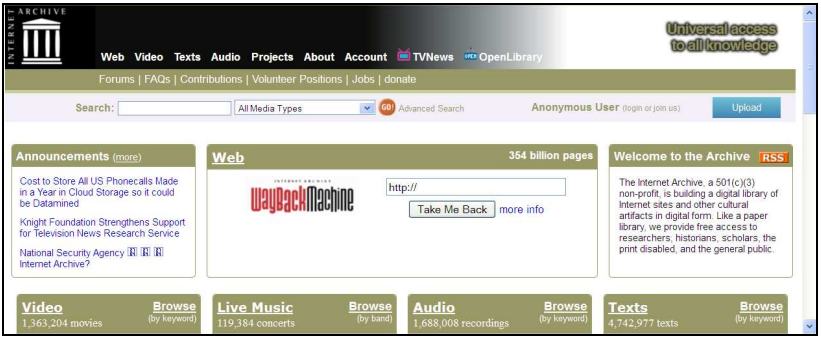


Рис. 9 – Заглавная страница веб-сайта www.archive.org

Одно из наиболее интересных практических применений этой технологии – восстановление документов, некогда опубликованных в веб-пространстве, но впоследствии удаленных. При этом рост глубинного веб грозит серьезными проблемами полноты в хранилище системы, связанными с

увеличивающимся количеством сайтов, эксплуатирующих различные технологии управления контентом, динамической публикацией документов из баз данных и т.п.

Сервисы работы с глубинным веб

Традиционные поисковые системы стремятся сузить пространство глубинного веб, постепенно захватывая такие ниши, как блоги, научные сайты, информационные агентства. Так, в качестве вспомогательных сервисов для поиска по глубинному веб от Google можно рекомендовать: Google Book Search (books.google.com) – поиск книг, Google Scholar (scholar.google.com) – поиск научных публикаций, Google Code Search (code.google.com) – поиск программного кода.

Система Goldfire Research от компании Invention Machine Corp. (inventionmachine.com) позволяет обрабатывать контент глубинного веб, размещенный на более чем 2000 сайтов правительственных, академических, исследовательских и коммерческих организаций США. Система Goldfire Research обладает информацией о механизмах доступа к базам данных глубинного веб и автоматически генерирует запросы к ним.

Исследовательская поисковая система Infovell из Калифорнийского университета Беркли (www.infovell.com) позволяет искать в глубинном веб по «ключевым фразам», от параграфов до целых документов, или даже наборам документов общим объемом до 25 тысяч слов. Система Infovell не зависит от языка, пользователи могут искать страницы на английском, арабском, китайском языках или же вводить в строке поиска математические уравнения, химические формулы.

Российская компания «Р-Техно» создала систему «it2b.интернетошпионопаук 3000+», предназначенную для выгрузки данных из невидимого сегмента сети Интернет. На основе этой системы построен поисковый

сервис Web Insight (www.r-techno.com/rtechno/online-services/ webinsight), обеспечивающий поиск по официальным сайтам и базам данных России и ближнего зарубежья, а именно, по документам Федеральной налоговой службы (ФНС), Федеральной службы судебных приставов (ФССП), Пенсионного фонда, Федеральной антимонопольной службы (ФАС), Трудовой инспекции, Федеральной регистрационной службы (ФРС), Министерства Российской Федерации по делам гражданской обороны, чрезвычайным ситуациям и ликвидации последствий стихийных бедствий (МЧС), Арбитражного суда, Министерства внутренних дел (МВД), Федеральной службы безопасности (ФСБ). Известны также такие базы данных службы «Р-Техно», как «Розыск Интерпола»; «Компании США уличенные в мошенничестве»; «Недобросовестные поставщики ФАС», «Должники металлургической отрасли» и т.п.

Существующие средства анализа и продвижения веб-ресурсов позволяют по-новому подойти к оценке соотношения объемов видимого и глубинного веб. Так на веб-сайте www.Cy-PR.com приводится информация о реальном количестве документов на исследуемом веб-сайте, представленном в RUNet, и о количестве документов, заиндексированных различными поисковыми системами, в том числе, Google и Яндекс. Получив репрезентативную выборку по сайтам, например, по рейтингу Рамблера top100 (top100.rambler.ru), можно получить оценку соотношения видимой и глубинной части в RUNet-сегменте веб-пространства.

Как показывают расчеты, объем информации, оказавшейся в глубинной части веб-пространства, превышает объем информации из видимой части примерно в 3-5 раз. Оказывается, за редким исключением, что чем крупнее ресурс, тем большая его часть относится к глубинному веб. В этом смысле небольшие веб-ресурсы выигрывают в доступности. Так как большая доля новостных документов оказывается в глубинном веб, то для задач бизнес-аналитики требуются специальные сервисы доступа к такой

информации. Именно такой сервис предоставляют службы интеграции новостного контента – архивы сетевых СМИ. Российские и украинские бизнес-аналитики активно используют крупнейшие архивы информации из открытых источников «Интегрум» (integrum.ru) и InfoStream (www.infostream.ua). Именно использование открытых источников позволяет конкурентной разведке действовать в рамках правового поля, но при этом иметь высокую эффективность.

Можно констатировать, что чем быстрее растет веб-пространство, тем хуже оно охватывается традиционными каталогами и поисковыми машинами. Из-за роста количества веб-сайтов и порталов, в которых применяются базы данных и динамические системы управления контентом, появления новых версий форматов представления информации, глубинный веб развивается очень интенсивно. С одной стороны, Интернет как огромное хранилище увеличивает объем информации, доступной «в принципе», но с другой стороны – растет информационный хаос, увеличивается энтропия сетевого информационного пространства. Все меньшая часть информационных ресурсов становится доступной пользователям реально.

Ведущие поисковые системы по-прежнему пытаются найти технические возможности для индексации содержимого баз данных и доступа к закрытым веб-сайтам, однако, их задачи объективно расходятся с задачами бизнес-аналитиков – ориентация традиционных поисковых служб на массовый сервис в данном случае оправдана. Таким образом, ниша для систем поиска в глубинном веб становится все шире.

1.3.3. Специальные базы данных

Как правило, для успешной аналитической деятельности, в частности, конкурентной разведки, должен быть создан и поддерживаться банк данных, включающий следующие основные базы данных:

- Конкуренты (действующие и потенциальные);

- Информация о рынке (тенденции, номенклатурная, ценовая, адресная информация);
- Технологии (продукты, выставки, конференции, ГОСТы, качество);
- Ресурсы (сырье, человеческие и информационные ресурсы);
- Законодательство (международные, центральные, региональные и ведомственные нормативно-правовые акты);
- Общие тенденции (политика, экономика, региональные особенности, социология, демография).

Если доступ к обычным интернет–ресурсам сегодня можно считать условно бесплатным, то, в большинстве случаев, доступ к базам данным требует не только регистрации но и оплаты таких услуг. Кроме того практически все они могут быть отнесены к так называемому «скрытому» веб-пространству.

Очень популярны среди специалистов по конкурентной разведке базы данных таможенных, налоговых и статистических органов, органов юстиции и судов, торгово-промышленных палат, органов приватизации и фондовых рынков, информационных, рейтинговых, аналитических и других агентств и т.д. Большую пользу приносят и отдельные доступные базы данных других контролирующих органов и организаций.

Традиционно конкурентная разведка опирается на такие источники информации, как опубликованные документы открытого доступа, которые содержат обзоры товарного рынка, информацию о новых технологиях, создании партнерств, слияниях и приобретениях, объявлениях о рабочих вакансиях, о выставках и конференциях, и т.п. Поэтому в последнее время все более популярны базы данных на основе архивов СМИ, в том числе и сетевых.

В «Большую тройку» мировых служб, занимающихся предоставлением пользователям доступа к деловой и аналитической информации, входят *LexisNexis, Factiva* и *Internet Securities.*

Крупнейшая в мире полнотекстовая онлайновая информационная система *LexisNexis* (www.lexisnexis.com), которая содержит свыше 2 миллиардов документов из 45 тыс. источников с архивом глубиной более 30 лет по бизнес-информации и более 200 лет по правовой информации, относится к разряду «скрытого» веб (рис. 10). Каждую неделю в архивы добавляется еще 14 млн. документов. В отличие от неструктурированных массивов «поверхностного» веб, пользователи ИС LexisNexis могут использовать мощные инструменты поиска для получения достоверной и классифицированной информации.

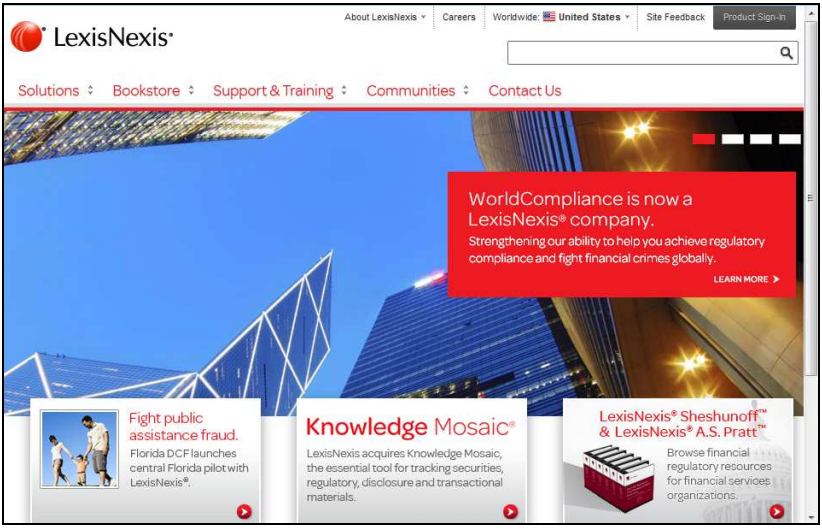


Рис. 10 – Фрагмент веб-сайта службы LexisNexis

Служба *Factiva* (global.factiva.com), подразделение компании Dow Jones, в настоящее время принадлежит компании News Corporation, занимается

предоставлением доступа к деловой и аналитической информации. В основе службы Factiva имеется более 35 тыс. первичных источников из 159 стран мира. В базе данных службы Factiva содержатся материалы более чем по 36,5 млн. компаний, а также полная подборка информации Investext.

Компания *Internet Securities* (www.internetsec.com), брэнд ISI Emerging Markets, охватывает 80 тематических информационных разделов, формируемых из 16 тыс. источников информации – тексты статей, финансовые и аналитические отчеты, корпоративная информация, макроэкономическая статистика, данные по рынкам (рис. 11). Основные продукты ISI Emerging Markets: CEIC Data, Emerging Market Information Service (EMIS), Islamic Finance Information Service (IFIS), IntelliNews, ISI Compliance Edition, ISI DealWatch.

Рис. 11 – Фрагмент веб-сайта службы Internet Securities

В России большой популярностью пользуются такие крупнейшие службы, как «*Интегрум*» (более 10 тыс. источников, сервисы «Анализ СМИ», «Архив СМИ»,

«Лента СМИ», БД «Компании», «Связи»), «*Медиология*» (13 тыс. источников: СМИ, телевидение, радио, газеты, журналы, информагентства, интернет, блоги, база данных из 30 тыс. объектов: компаний, персон и брендов), «*Яндекс.Новости*» (служба автоматической обработки и систематизации новостей, свыше 4000 источников, не допускается участие в работе службы материалов, содержащих сообщения неновостного характера), Public.Ru – крупная онлайн библиотека русскоязычных СМИ. С 2000 года *Public.Ru* создает свою базу данных, которая хранит архивные материалы российских изданий с 1990 года. В архивах базы данных доступно более 70 млн. статей русскоязычных СМИ собранных из 4600 источников. Основные виды источников, содержащихся в базе данных: федеральные издания; региональные издания; информационные агентства; телеканалы; радиостанции; интернет-издания.

В Украине эту нишу занимает система контент-мониторинга интернет-СМИ *InfoStream* (свыше 6 тыс. источников информации, более 100 миллионов документов в архиве).

Украинская корпорация «*Медіа-простір*» (550 источников информации, 25 региональных информационных бюро), которая осуществляет анализ информационного пространства Украины, обеспечивает предоставление медиа-обзоров. Систематизация информационных сообщений происходит по четырем объектам привязки: предметные сферы; политические субъекты; личности; территории.

Информационно-мониторинговая система *Web-Observer* в «базовой комплектации» охватывает 500 источников. Система внедрена в информационном агентстве УНИАН, на ней базируется сервис «УНИАН-монитор».

Украинская система интернет-мониторинга *MonitorIX* обеспечивает мониторинг интернет-источников (как сетевая информационно-поисковая система), СМИ (130 изданий), ТВ, блогов и форумов.

Предоставляет клиентам результаты оперативного и архивного мониторинга.

Приведем еще один пример зарубежной базы данных из «скрытого» веб. Корпорация LexisNexis предоставляет сервис *Auto TrackXP*, вошедший в список двадцати крупнейших «скрытых» веб-сайтов мира (по рейтингу BrightPlanet). Auto TrackXP представляет собой базу данных объемом 30 Терабайт (ТБ), охватывающую практически все аспекты гражданской жизни США. База данных Auto TrackXP содержит информацию практически о каждом гражданине США. TestProfiles.com – часть ChoicePoint Online – содержит личные характеристики и сведения о компетентности граждан США.

Система широко используется как легальный ресурс для задач конкурентной разведки. Вместе с тем, сегодня американцы повсеместно выражают возмущение, обнаруживая существование подобных сервисов, видя в этом нарушение своих гражданских прав.

Сервис *Insight Profiles* (www.insightprofiles.com) содержит характеристики и сведения о способностях и компетентности граждан США. В частности, чтобы определить, не завладел ли человек чужими документами, в рамках сервиса *Insight Profiles* организован дополнительный платный сервис *ProCheck* (procheck.com), позволяющий сопоставлять информацию из различных источников и государственных каталогов.

Для частных любителей составления «досье» *ChoicePoint* предлагает более скромный, но не менее любопытный набор сервисов (www.choicetrust.com). Подозрительные пациенты с помощью *Doctor Check* имеют возможность самостоятельно выбрать или проверить квалификацию врачей 40 различных специализаций. Отчет, получаемый с помощью системы, может, например, служить для страховой компании поводом в отказе выдачи полиса.

В России и Украине популярны такие базы данных, как:

«*Лабиринт*» (www.labyrinth.ru) – база данных, составленная на основе публикаций ведущих бизнес-изданий, предназначена для помощи при выполнении аналитических и исследовательских работ, написании статей, комментариев, докладных записок, пр. В базах данных представлены биографии российских деятелей, справки об организациях и компаниях, информация о субъектах Российской Федерации и другой справочный материал;

«Компасс» (www.kompass.com/ru) – база данных, позиционирующая себя как международную информационную B2B (типа «бизнес для бизнеса») поисковую систему, обеспечивает поиск по компаниям, товарам и услугам, руководителям с целью формирования баз данных целевого маркетинга и сбыта, потенциальных клиентов;

«KAPE» (kare.pulscen.com.ua) – базы данных предприятий Украины – 384000 компаний и агропромышленного комплекса Украины – 218000 компаний;

База данных *Dun & Bradstreet* (*D&B*). Национальные представительства *D&B* в Украине – это компания «Бизнес-мониторинг», которая входит в состав группы компаний «Авеста-Украина» и представляют направление деловой информации. Компания обеспечивает подключение к Системе Профессионального Анализа Рынков и Компаний (СПАРК-Россия) и к базе данных Dun & Bradstreet (D&B). В России подключение к этой базе данных обеспечивает информационное агентство Интерфакс (www.dnb.ru);

Базы данных международной корпорации *Creditreform*, представленной в России компанией «Кредитреформ РУС» (www.creditreform-rus.ru), а в Украине информационным агентством «Кредитреформ Украина» (www.creditreform.ua). Компании обеспечивают

доступ к международной сети содействия информационному бизнесу BIGNet (Business Information Group Network). Эта сеть объединяет независимые организации, которые предоставляют бизнес-справки по всему миру в режиме онлайн (свыше 8 млн. в год), а также доступ к базе данных BIGNet, а также к собственной базе данных (www.crefoport.ru), содержащей данные о 30 млн. компаний.

Europages (www.europages.eu) – Европейская бизнес-директория – информационно-поисковая B2B-система, охватывающая свыше 2 млн. поставщиков, производителей и дистрибьютеров в Европе и во всем мире.

Задача полного перечисления всех источников информации практически невыполнима, так как этот рынок очень динамичен, постоянно появляются новые базы данных, происходит слияние существующих источников, поглощение слабых сильными. Вместе с тем, одно из правил конкурентной разведки формулируется таким образом: «чем большим количеством независимых источников подтверждается информация – тем более она достоверна».

Наряду с базами данных, одним из самых эффективных источников информации могут служить отчеты и справки аутсорсинговых компаний, профессионально занимающихся конкурентной разведкой и сбором сведений о коммерческих структурах и рынках. Их продукция, на самом деле, и является результатом конкурентной разведки.

В мире существует множество таких специальных компаний. Одной из таких крупнейших компаний, которой принадлежит около 80% западного рынка, является американская компания, *Dun & Bradstreet (D&B)*, чья база данных упоминалась нами выше. Справка по любой компании в этой службе оценивается из расчета в среднем 100 долларов и выше. Более серьезный анализ рынка или конкурента может обойтись и в 10 тыс. долларов. Сроки исполнения – от нескольких часов (информация присутствует в базе

данных) – до нескольких суток для справок и до нескольких месяцев для аналитической работы.

На европейском рынке не менее известны названная выше ирландская компания *Creditreform,* немецкая *Schufa Holding AG* (479 млн. документов в базах данных, в том числе, 66 млн. о физических лицах), австрийская *Intercredit Infromation Holding*, латвийская *Coface IGK* (известна IGK System – *база данных* должников, включающая сведения о текущих долгах, судебных исках, а также процессах неплатежеспособности) и многие другие. Некоторые из этих компаний совмещают функции конкурентной разведки с другими видами деятельности, например, обязанностями кредитных бюро.

Общей проблемой при обращении за информационными справками в западные агентства, имеющие представительства в России и Украине, является то, что, как правило, информация, предоставляемая в отношении западных нерезидентов, намного обширнее и качественнее, чем та, что предоставляется в отношении отечественных фирм. В связи с чем, в таких случаях целесообразно обращаться к местным информационным компаниям, результаты оказываются дешевле и качественнее.

На российском рынке в сфере конкурентной разведки пользуются популярностью информационные отчеты компаний *«Р-Техно», «Медиология», «Специальная Информационная Служба», «Интегрум», «Кронос-Информ»* и многих других.

В Украине также существует целый ряд подобных компаний, среди которых можно назвать «*Авеста-Украина», «СИДКОН», Межбанковская служба безопасности «СКИФ»* и другие.

Все отечественные и зарубежные информационные компании имеют свои представительства и принимают заказы в Интернете, в связи с чем их можно отнести к специфическим интернет-источникам.

Следует также отметить, что, несмотря на то, что в случае заказа услуг аутсорсинговой компании, она делает большую часть информационной работы за клиента, окончательные выводы и решения, рекомендации для принятия управленческих решений все-таки остаются за ним. Только клиент может обладать всей необходимой полнотой внешней и инсайдерской информации.

1.3.4. Социальные сети

Понятие социальных медиа

Социальные медиа представляют собой совокупность онлайн-сервисов и интернет-приложений, которые позволяют пользователям общаться друг с другом в том числе и в режиме реального времени. При этом пользователи могут обмениваться между собой мнениями, новостями, информацией, в том числе и мультимедийной. Социальные медиа базируются на идеологической и технологической базе веб 2.0, позволяющих создание и обмен контентом, созданным самими пользователями (User-Generated Content), в отличие от предшествующей концепции веба, предполагающей, как и в случае традиционных СМИ, централизованное создание контента, поставляемого пользователям-читателям.

Очевидно, социальные медиа являются самым ценным источником информации для конкурентной разведки, предоставляя абсолютно на легальных условиях разностороннюю информацию о людях, событиях, компаниях, брендах, продуктах. Получившие в последнее время широкое распространение такие явления, как информационные операции, активное информационное противодействие в рамках конкурентной борьбы, сетевая мобилизация, во многих случаях базируются на манипулировании данными именно в социальных медиа.

Выделяют семь разновидностей социальных медиа, это: социальные сети; блоги; форумы; сайты отзывов; серверы фото- и видеохостинга; виртуальные

службы знакомств и геосоциальные сети. Следует отметить, что четкие границы между этими разновидностями размыты.

Под социальной сетью в сети Интернет (social networking service) понимается онлайн-сервис, предназначенный для построения, отображения и организации социальных взаимоотношений, обеспечивающий предоставление широкого спектра возможностей для обмена информацией, возможность пользователя предоставить информацию о самом себе (создать свой профиль), построить связи, найти друзей по интересам, подключить родственников, коллег, одноклассников и т. п.

Под блогом (blog, от web log) понимают веб-сайт, основное содержание которого – это периодически добавляемые пользователями записи (текст, изображения или мультимедиа). Для блогов характерны недлинные записи (особенно, в случаях так называемых «микроблогов») временной значимости, блоги обычно публичны и предполагают сторонних читателей, которые могут вступить в публичную полемику с автором (в комментарии к блогозаписи или своих блогах). Совокупность всех блогов в сети Интернет называют блогосферой.

Веб-форумы представляют собой веб-приложения, предназначенные для организации общения посетителей некоторых интернет-ресурсов (веб-сайтов или порталов). На ресурсах веб-форума пользователи задают интересующие их темы, которые затем обсуждаются и другими пользователями путем размещения сообщений (постинга) внутри этих тем.

Веб-сайты отзывов создаются с целью повышения эффективности и качества предоставляемых (не обязательно в интернет-среде) услуг и товаров. Пользователи, посещая веб-сайты отзывов, оставляют там свои сообщения, участвуют в анкетированиях, формируют мнения о той или иной услуге или товаре.

Фотохостинг (photo hosting) – это веб-сайт, позволяющий публиковать любые изображения (чаще всего, цифровые фотографии) в сети Интернет. Основное преимущество фотохостинга – удобство демонстрации размещенных фотографий. Соответственно, видеохостинг – это веб-сайт, позволяющий загружать и просматривать видеоинформацию в веб-браузере. Видеохостинг набирает популярность в связи с развитием широкополосного доступа в Интернет.

Виртуальная служба знакомств представляет собой интернет-сервис, оказывающий услуги по виртуальному знакомству пользователей с целями общения, создания семьи, серьезных отношений и др. При использовании виртуальной службы знакомств пользователь создает анкету, в которой указывает свой псевдоним (никнейм) и другие параметры, запрашиваемые службой (пол, возраст, цель знакомства, интересы, фотографии). После регистрации пользователь может общаться с другими пользователями, получать сообщения и отвечать на них.

Геосоциальные сети (GeoSocial Network) – это разновидность социальных сетей, в которых пользователи оставляют данные о своем местонахождении, что позволяет объединять и координировать их действия на основании информации о том, какие люди присутствуют в тех или иных местах, какие события происходят в этих местах.

Термин «социальная сеть» обозначает сосредоточение социальных объектов, которые можно рассматривать как сеть (или граф), узлы которой – объекты, а связи – социальные отношения. Этот термин был введен в 1954 году социологом из «Манчестерской школы» Дж. Барнсом (J. Barnes) в работе «Классы и сборы в норвежском островном приходе». Во второй половине XX-го столетия понятие «социальная сеть» стало популярным у западных исследователей, при этом как узлы социальных сетей стали рассматривать не только представителей социума, но и другие объекты, которым присущий социальные связи. Сегодня термин

«социальная сеть» обозначает понятие, оказавшееся шире своего социального аспекта, оно включает, например, многие информационные сети, в том числе и WWW. Рассматривают не только статистические, но и динамические сети, для понимания структуры которых необходим учет принципов их эволюции.

Сегодня под термином «социальные сети» (*Social Networks*) понимают, прежде всего, онлайн-сервисы в сети Интернет, предназначенные для формирования, отображения и упорядочения социальных взаимоотношений. Выделяют следующие особенности социальных сетей:

- предоставление пользователям широкого спектра возможностей для обмена информацией;
- создание профилей пользователей, в которых требуется указать некоторое количество персональной информации;
- друзьями в социальной сети становятся преимущественно не виртуальные, а реальные друзья.

Социальные сети предоставляют такие возможности:

- активного общения;
- создания публичного или закрытого профиля (Profile) пользователя, содержащего персональные даные;
- организации и ведения пользователем списка других пользователей, с которыми у него имеются некоторые социальные отношения;
- просмотра связей между пользователями внутри социальной сети;
- образования групп пользователей по интересам;
- управления содержимым в рамках своего профиля;

- синдикации контента;
- подключения различных приложений.

Ресурсы социальных медиа

В список крупнейших социальных сетей, которые могут быть интересными для конкурентной разведки, можно включить:

- Facebook;
- ВКонтакте;
- Одноклассники;
- Google+;
- Мой Круг;
- LinkedIn;
- Badoo;
- Livejournal;
- Twitter.

Facebook (www.facebook.com) – крупнейшая социальная сеть, основанная в 2004 году М. Цукербергом и его компаньонами. Начиная с сентября 2006 года социальная сеть доступна для пользователей сети Интернет. По данным на октябрь 2012 года аудитория *Facebook* составила 1 миллиард пользователей. Суточная активная аудитория превышает 525 миллионов человек. Около 500 млн человек в месяц используют мобильные приложения *Facebook*. Каждый день в социальной сети пользователи оставляют 3,2 миллиарда «лайков» и комментариев и публикуют 300 миллионов фотографий. На сайте зафиксировано 125 миллиардов «дружеских связей». Ежемесячное количество просмотров страниц *Facebook* превышает 1 триллион.

«ВКонтакте» (vk.com) – крупнейшая в Рунете социальная сеть, созданная П. Дуровым в 2006 г., позиционирующая себя как «современный, быстрый и

эстетичный способ общения в сети». По данным на начало 2013 года ежедневная аудитория «ВКонтакте» составляет более 43 миллионов пользователей. Пользователям «ВКонтакте» доступен характерный для многих социальных сетей набор возможностей: создавать профиль с информацией о себе, производить и распространять контент, управлять настройками доступа, взаимодействовать с другими пользователями приватно и публично, отслеживать через ленту новостей активность друзей и сообществ.

Социальная сеть «*ВКонтакте*» предлагает сторонним ресурсам использовать специально разработанные инструменты — виджеты — для глубокой интеграции с социальной сетью. Эти решения позволяют встраивать в веб-сайты систему комментариев для пользователей, сообщества, системы опросов, а также возможность легко поделиться ссылкой на материал с другими пользователями и авторизоваться на сайте. У сайта имеется мобильная версия, расположенная по адресу m.vk.com. 24 мая 2013 г. Роскомнадзор внёс домен vk.com и его IP-адрес в «Единый реестр запрещённых сайтов», однако уже через несколько часов удалил его оттуда, обосновав ошибку человеческим фактором.

Одноклассники (http://www.odnoklassniki.ru/) – мультиязычная социальная сеть, используемая для поиска одноклассников, однокурсников, бывших выпускников, а также родных и близких родственников и общения с ними. Проект запущен в 2006 г. На начало 2013 г. количество пользователей сети превысило 200 млн, а посещаемость – более 44 мил. посетителей в сутки. С 1 августа 2014 года в РФ вступил в силу закон, который дает право Федеральной службе безопасности получать все личные данные пользователей российских интернет-ресурсов. Поскольку «Одноклассники» находится на территории России, закон распространяется на пользователей этой социальной сети.

Google+ (plus.google.com) – социальная сеть от компании Google, официально начавшая свою работу в 2011 г. На начало 2012 г. количество зарегистрированных в Google+ пользователей превышало девяносто миллионов человек. Сервис предоставляет возможность общения через Интернет с помощью специальных компонентов: *Круги*, *Темы*, *ВидеоВстречи*, *Мобильная версия*. Основополагающими принципами действия сервиса являются пользователи, приватность и живое общение. Информация, которой делятся участники сети, влияет на персонализированные результаты поиска Google. В основе работы Google+ лежит концепция кругов (*Circles*), благодаря которым пользователь регулирует своё общение. На основе кругов пользователь делится контентом, определяя, какой круг будет иметь доступ к информации, а какой нет. Обмен пользовательскими материалами идёт в специальной ленте (*Stream*), в которой можно следить за обновлениями участников кругов. Google была представлена также и мобильная версия социальной сети, в которой есть две уникальных функции: мгновенная загрузка фото и Чат (*Huddle*). Социальная сеть Google+ позволяет получать хорошие позиции в поиске Google.

«Мой Круг» (www.moikrug.ru) – русскоязычная социальная сеть, направленная на установление деловых контактов между людьми. Архитектура сети представляет собой круги пользователей, где первый круг – это близкие друзья пользователя, которым он доверяет свою контактную информацию, второй круг – это друзья друзей пользователя, а третий соответственно друзья друзей его друзей. Сеть «Мой Круг» была создана в 2005 г. В 2007 г. проект был куплен компанией «Яндекс». Теперь он является одним из сервисов Яндекса. В 2009 году к сервису был подключен экспорт вакансий из крупнейших тематических сайтов: hh.ru и rabota.mail.ru. В 2011 году социальная сеть «Мой круг» завершила процесс интеграции с социальными платформами – появилась возможность поиска друзей

через аккаунты Facebook, Twitter, LiveJournal и LinkedIn.

LinkedIn (www.linkedin.com) – социальная сеть для поиска и установления деловых контактов. В социальной сети LinkedIn зарегистрировано свыше 200 миллионов пользователей из 200 стран, представляющих 150 отраслей бизнеса. Социальная сеть LinkedIn, основанная Р. Хоффманом, была запущена в эксплуатацию в 2003 г. Эта социальная сеть предоставляет возможность зарегистрированным пользователям создавать и поддерживать список деловых контактов. Контакты могут быть приглашены как из сайта, так и извне, однако LinkedIn требует предварительного знакомства с контактами. В случае, когда пользователь не имеет прямой связи с контактом, он может быть представлен через другой контакт. Список контактов LinkedIn может использоваться для: расширения связей, поиска компаний, людей и групп по интересам, публикации резюме и поиска работы, а также рекомендации одних пользователей другими, публикации вакансий, создания групп по интересам. Социальная сеть LinkedIn также позволяет публиковать информацию о деловых поездках и конференциях.

Badoo (badoo.com) – социальная сеть знакомств, основанная в 2006 г. российским бизнесменом А. Андреевым. По состоянию на 2013 г. в социальной сети *Badoo* зарегистрировано более 180 миллионов пользователей. Зарегистрировавшись, пользователь может общаться в чате, загружать на сайт свои фотографии, связываться с друзьями в своем регионе или за его пределами. Существуют также премиум-услуги, которые являются платными. Они предоставляются тем, кто хочет иметь большую популярность, расширить круг знакомств. За время своего существования компания Badoo выпустила несколько продуктов под свободной лицензией, включая различные улучшения языка программирования PHP, сервер Pinba, собирающий статистику в реальном времени, бесплатный быстрый шаблонизатор Blitz для PHP.

«Живой Журнал», ЖЖ, LiveJournal, LJ (www.livejournal.com) – платформа для ведения онлайн-дневников (блогов), созданная в 1999 г. американским программистом Б. Фицпатриком. *«Живой Журнал»* предоставляет пользователям возможность публиковать свои и комментировать чужие записи, вести коллективные блоги («сообщества»), добавлять в друзья («френдить») других пользователей и следить за их записями в «ленте друзей» («френдленте»). По данным статистики LiveJournal.com на конец 2012 г. в «*Живом Журнале*» зарегистрировано более 40 млн пользователей. Среди настроек, функций и опций «*Живого Журнала*» следует выделить: разные типы записей и возможности их комментирования; указание расширенных сведений о пользователе; друзья и лента друзей; картинки пользователей; функции безопасности аккаунта.

Твиттер (англ. *Twitter* – «щебетать») (twitter.com) – сервис, позволяющий пользователям отправлять короткие текстовые заметки (до 140 символов), используя веб-интерфейс, SMS, средства мгновенного обмена сообщениями или сторонние программы-клиенты. Созданный Дж. Дорси в 2006 г., по состоянию на начало 2011 года сервис насчитывает свыше 200 млн. пользователей, из них 50 млн. пользуются Твиттером ежедневно. Порядка 55 % пользуются Твиттером на мобильных гаджетах. Особенностью Твиттера является публичная доступность размещённых сообщений; это называется микроблоггингом.

Мониторинг социальных медиа

В последнее время выделилось отдельное научное направление – анализ социальных сетей (SNA, Social Networks Analysis), которое базируется, с одной стороны, на социологии, а с другой на теории сложных сетей (Complex Networks) [Newman, 2003].

Благодаря мониторингу социальных сетей (SMM – Social Network Monitoring) бизнес-аналитики значительно увеличили возможность лучшего понимания объектов исследования. В настоящее время появился значительный спрос на системы и сервисные

решения, которые позволили бы анализировать мнения, которые появляются в социальных сетях. Сегодня бурно формируется рынок SMM, охватывающий как бесплатные сервисы (Google Alerts, Яндекс блоги), так и целый ряд коммерческих служб (Radian6, Brand24, YouScan), которые предоставляют качественный сервис мониторинга социальных сетей.

В настоящее время системы автоматического мониторинга социальных медиа получают большую популярность, это связано с приходом политики и бизнеса в социальные сети и с усложнением задач, которые необходимо решать в данной среде.

Среди основных задач, которые решаются с помощью систем автоматического мониторинга социальных медиа, можно выделить следующие.

- Обнаружение негативной информации

Социальные сети обеспечивают высокую скорость распространения информации, в том числе негативных, критических высказываний. Для некоторых видов деятельности, таких как страхование или банковское дело, такая информация, которая за небольшое время может превратиться в так называемый «информационный взрыв», резкое возрастание напряженности.

Следовательно, нейтрализация негативной информации в социальных сетях и блогосфере в кратчайшие сроки является важнейшей задачей, от решения которой может зависеть судьба всего бизнеса. Для решения этой задачи применяются системы оперативного мониторинга социальных медиа.

- Анализ состояния конкурентов

Анализ мнений участников социальных сетей и блогосферы о конкурентах, учет откликов, упоминаний, тональности сообщений, изучение информации, которую конкуренты размещают о себе, о своей деятельности, продукции, ценах или их маркетинговой политики являются важными задачами маркетинга.

Эффективная система мониторинга социальных медиа должна обеспечить автоматизацию сбора, систематизации и анализа информации о конкурентах, обеспечить возможность сравнения отдельных показателей различных конкурентов.

- Оценка эффективности рекламных кампаний

Важная функция систем мониторинга – оценка эффективности проводимых рекламных кампаний как в традиционной среде, так и в сети Интернет, в социальных сетях. Ретроспективный анализ данных в этом случае позволяет отслеживать эволюцию общественного мнения по отношению к компании, продукции, бренду.

- Получение обратной связи

Обратная связь с клиентами позволяет компании обращать внимание на важные для клиентов аспекты (сильные и слабые стороны продукции, сравнение с конкурентами, аналогичной или заменяющей продукцией), при необходимости корректировать свою позицию на рынке, рекламную и маркетинговую политику

- Участие в тематических дискуссиях

Для донесения до клиентов информации о компании, продукции, бренде, конкурентных преимуществах, нейтрализации негативной информации необходимо участие в форумах, тематических дискуссиях. Эффективная система мониторинга социальных медиа должна обеспечивать оперативное своевременное нахождение таких форумов, обсуждений, дискуссий.

Эффективный мониторинг социальных медиа обеспечивает следующие преимущества.

- Непрерывность анализа. Мониторинг социальных медиа ведется непрерывно, отчет по состоянию объекта мониторинга можно получить в любое время.

- Сохранение запроса на мониторинг. Система мониторинга социальных медиа обеспечивает сканирование сетей в соответствии с сохраненными запросами пользователей, которые могут корректироваться при необходимости.
- Ретроспективность мониторинга. Общедоступные бесплатные системы мониторинга социальных медиа зачастую не предусматривают возможность просмотра большого количества страниц результатов ретроспективного поиска. Специальные системы мониторинга социальных медиа, как правило, позволяют обходить это ограничение.
- Автоматическое обобщение, систематизация полученных данных, при необходимости занесение их в соответствующие формы, таблицы.
- Аналитические функции. Автоматические системы сами способны анализировать данные, в частности, выявлять информационные дубли, формировать краткие резюме, классифицировать документы, определять их тональность.
- Автоматическое формирование отчетов. Современные системы мониторинга включают функции генерации отчетов в виде графиков и диаграмм, позволяющих наглядно демонстрировать динамику тематических информационных потоков, тональность соответствующих сообщений.

Одна из наиболее эффективных архитектур организации систем мониторинга социальных медиа трехуровневая.

Уровень 1: перманентный мониторинг

Перманентный мониторинг предназначен для обеспечения немедленной реакции пользователя на некоторые сообщения в сетях, например, на негативные

упоминания компании, бренда или продукта, или на срочные вопросы клиентов компании. Необходимость постоянного просмотра упоминаний обусловливает мониторинг в режиме онлайн или хотя-бы несколько раз в сутки. При этом, как правило, не требуется использования развитых аналитических возможностей. Кроме того, существуют автоматические системы, рассылающие уведомления при появлении публикаций, соответствующих определенным словосочетаниям.

Уровень 2: периодический мониторинг трендов

Цель мониторинга второго уровня – анализ тенденций в информационном пространстве за определенный период. На втором уровне анализируются:

- Тренды и эмоциональная окраска упоминаний, информационные поводы, события, за прошедший период.
- Динамика информационных потоков, выражающаяся в том, насколько изменилось количество упоминаний объекта мониторинга за прошедший период, насколько изменилось соотношение эмоциональной окраски упоминаний.
- Получение ответов на актуальные для пользователя вопросы, задаваемые клиентам (как реальными, так и потенциальными). При этом как вопросы, так и обобщенные ответы должны уточняться за счет получения новых данных.

Уровень 3: стратегический мониторинг

Цель третьего уровня мониторинга – охват и формирование понимания места объекта мониторинга в информационном пространстве, выявление и обоснование основных тенденций, закономерностей, возможностей и опасностей. Для этого анализируется весь информационный поток мониторинга вне

временного ограничения. На третьем уровне анализируются:

- Информационные корреляции между событиями и процессами, между информационными поводами и реакциями пользователей.
- Целевые ресурсы, на которых встречается наибольшее количество упоминаний объекта мониторинга, лояльных пользователей, мест информационной напряженности.
- Влияние отдельных публикаций (позитивных или негативных упоминаний объекта мониторинга) на его реальное состояние.
- Резонанс публикаций в социальных медиа, общий охват упоминаний компании, среднее количество републикаций сообщения, количество вовлеченных в обсуждение пользователей сетей.
- Выявление лидеров мнений, как лояльных, так и тех, кто упоминает объект мониторинга в отрицательном ключе, агрессоров — пользователей, которые часто публикуют негативную информацию.

Эффективные системы мониторинга социальных сетей позволяют значительно снизить трудозатраты за счет автоматизации рутинных процессов и достичь высокой точности за счет систематизации данных и применения аналитического инструментария, помогающего делать выводы о том, как меняется ситуация, связанная с объектом мониторинга, в компании в целом.

Существующие системы монитоинга социальных медиа

Лидером мирового рынка мониторинга социальных медиа является канадский социальный аналитический сервис Radian6, который был приобретен Salesforce за $326 млн. в марте 2011. Сервис предлагает компаниям услугу мониторинга в режиме

real time. С помощью Radian6 компании могут узнать, что пользователи соцсетей говорят о них и их продукции в Facebook, Twitter, YouTube и LinkedIn, а также в блогах и на веб-форумах.

Еще один польский стартап Brand24 демонстрирует высокую динамику развития на рынке мониторинга social media. Число их клиентов переросло несколько сотен. Среди них Panasonic, Intel, IKEA, Raiffeisen Bank, AirFrance и другие крупные международные корпорации и бренды. Основные преимуществ Brand24 это скорость актуализации результатов поиска, широкий диапазон мониторинга, функциональность, а также usability для пользователей. Недавно стартап объявил о выходе на международный рынок, сделав сервис доступным на английском языке.

В настоящее время динамично развивающимся игроком рынка на постсоветском пространстве является компания *YouScan*, которая обеспечивает анализ социальных медиа, предоставляя пользователям удобный интерфейс, а также возможность наблюдать результаты мониторинга в режиме реального времени в виде инфографики.

Рассмотрим еще несколько сервисов для автоматического мониторинга социальных медиа, сосредоточив внимание на наиболее доступних. Среди других известных систем мониторинга социальных медиа можно назвать:

Seesmic (seesmic.com) – бесплатный сервис мониторинга социальных медиа. Поддерживает мониторинг таких ресурсов, как: Twitter, Facebook, LinkedIn, Chatter, Google Buzz, Ping.fm. Есть приложения как для веб, так и для персонального компьютера, iPhone, Android, Windows Mobile.

Socialmention (www.socialmention.com) – платформа бесплатного поиска и анализа информации в социальных сетях. Система ищет упоминания в выбранных сетях или во всех сетях сразу. Предоставляет анализ тональности упоминаний,

связанные ключевые слова, популярные источники и многое другое. Охват системы – более 100 социальных медиа, включая социальные сети, социальные закладки, блоги, форумы и многое другое.

Hootsuite (hootsuite.com) – многофункциональный сервис для работы с социальными медиа. Система Hootsuite позволяет работать с аккаунтами Twitter, Facebook, LinkedIn, MySpace и Foursquare, с блогами на WordPress. Сервис HootSuite является сертифицированным партнёром Twitter. Обеспечивает постинг (posting) по расписанию, возможность отслеживать сообщения по ключевым словам и упоминаниям. Система HootSuite также предоставляет полноценную интеграцию с Facebook. Система HootSuite является условно-платной, есть бесплатная версия (аналитика, 5 социальных профилей, 2 RSS/Atom ленты. Доступна на мобильных платформах: iPhone, Android, Blackberry. Все мобильные программы бесплатны.

S-monitor (s-monitor.com) – сервис онлайн-мониторинга социальных медиа и веб-ресурсов. Поддерживает мониторинг социальны сетей Facebook и ВКонтакте; блогов: LiveJournal и Twitter, нескольких тысяч веб-сайтов, представленных на различных языках. В рамках сервиса предоставляются отчеты по количеству сообщений с упоминаниями ключевых слов, брендов, топонимов, авторов, источников, тональности.

YouScan (www.youscan.ru) – сервис мониторинга русскоязычных социальных медиа. *YouScan* отслеживает упоминания в блогах, форумах, социальных сетях (Facebook, ВКонтакте), Twitter, YouTube, и предоставляет результаты мониторинга в аналитическом интерфейсе с функциями одновременной работы нескольких сотрудников.

BuzzLook (buzzlook.ru) – сервис мониторинга социальных медиа: «ВКонтакте», Facebook, Livejournal, Flickr, YouTube и Twitter, позволяющий: следить за репутацией бренда; изучать деятельность конкурентов в сети; отвечать на вопросы клиентов в социальных сетях;

собирать предложения от клиентов; поддерживать онлайн-сообщества.

IQBuzz (www.iqbuzz.ru) – сервис для мониторинга социальных медиа – большого количества источников и площадок, таких как LiveInternet, LiveJournal, Twitter, Яндекс.Блоги, сервисы видеохостинга RuTube и YouTube, различные новостные, развлекательные, специализированные, тематические и региональные порталы. Система обеспечивает круглосуточный мониторинг, позволяет получать информацию практически в режиме реального времени. Система IQBuzz позволяет определять тональность пользовательских сообщений, анализировать социально-демографические характеристики их авторов на основании информации из профайлов социальных сетей. Имеется возможность подключения по запросам пользователей новых источников для мониторинга.

Socialbakers (www.socialbakers.com) – сервис сбора статистики о работе социальных сетей, называющий себя «сердцем статистики Facebook». Система Socialbakers известна своими рейтингами брендов на Facebook в разных категориях. Кроме Facebook сервис Socialbakers предоставляет возможность бесплатного мониторинга информации в таких сетях, как в Twitter, Google+, LinkedIn.

SocialSeek (socialseek.com) – простой в использовании бесплатный сервис мониторинга нескольких социальных медиа в режиме реального времени. Обеспечивает поиск в новостях, блогах, Twitter, Facebook, Youtube.

Socialpointer (www.socialpointer.com) – простой сервис мониторинга в социальных сетях, новостях, блогах. Имеется базовая аналитика.

PeerIndex (www.peerindex.net) – бесплатный сервис анализа социальных медиа, прежде всего Twitter, Facebook, LinkedIn. Определяет размеры «социального капитала» или влиятельности компании, профессиональности, объема публикаций и др.

PostRank (www.postrank.com) – сервис компании Google, позволяющий в режиме реального времени анализировать данные по темам, тенденциям, событиям, имеющим отношение к личности или бизнесу.

Topsy (topsy.com) – бесплатный сервис поиска в режиме реального времени по социальным медиа.

HowsSciable (www.howsociable.com) – бесплатный инструмент мониторинга брендов и ключевых слов в 32 социальных сетях.

Twitalyzer (www.twitalyzer.com) – аналитическая программа-клиент для Твиттера, позволяющая отслеживать количество переходов, анализировать позитивные и негативные комментарии, сегментировать аудиторию. Интегрирована с системой Google Analytics, выводит интерактивные диаграммы и графические инструменты.

WildFire (monitor.wildfireapp.com) – многофункциональный онлайн-сервис для коммерческого медиа-маркетинга в социальных сетях, включающий инструмент Wildfire Messages, предназначенный для создания, мониторинга и управления сообщениями. Позволяет настроить отложенный постинг сообщений в социальные сети по расписанию. Предоставляет полноценный функционал для продвижения брендов в различных социальных сетях.

Kurrently (www.kurrently.com) – бесплатная поисковая система по социальным сетям Twitter и Facebook, позволяющая отслеживать, распространять целевую информацию по социальным сетям.

Trackur (www.trackur.com) – коммерческий онлайн- инструмент мониторинга и анализа социальных медиа. Позволяет отслеживать репутацию брендов по новостным веб-сайтам, блогам, форумам, социальным сетям Twitter, Google+ и Facebook.

Babkee (www.babkee.ru) – система мониторинга упоминаний в социальных медиа. Позволяет решать такие задачи, как оценка репутации бренда; анализ эффективности рекламных кампаний в сети Интернет; проведение маркетинговых исследований рынка, конкурентов и целевой аудитории; реагирование на обращения пользователей и их поддержка. Система позиционируется как уникальная услуга оценки значимости сообщений. Есть возможность бесплатного использования.

Buzzware (www.buzzware.ru) – сервис мониторинга социальных медиа, позволяющий исследовать мнения пользователей о брендах, которые они выражают в блогах и социальных сетях. Сервис можно использовать для репутационного анализа, изучения конкурентов, получения представлений о пользовательском опыте и ожиданиях и, конечно же, для оценки успешности проведенных кампаний в сети Интернет.

SemanticForce (www.semanticforce.net) – сервис, обеспечивающий мониторинг неструктурированных источников – комментариев в сетевых СМИ и интернет-магазинах (рис. 12). Выдает более 20 видов аналитических отчетов. Сервис SemanticForce интегрирован с внешними системами: Klout, Copiny, GoogleAnalytics.

Brandspotter (www.brandspotter.ru) – система мониторинга социальных медиа, предлагающая стандартный набор услуг, включающий определение эмоциональной окраски высказываний, получение статистики мониторинга сообщений по темам, платформам, авторам.

Крибрум (www.kribrum.ru) – система мониторинга социальных сетей, позволяющая отслеживать и анализировать упоминания бренда, продуктов или услуг, ключевых персон, событий, географических названий. Содержит инструментарий автоматического оценивания эмоциональной окраски сообщений и построения интерактивных отчетов.

Система сообщения, в которых бренд упоминается лишь вскользь. Данные начинают отображаться в системе через 2–4 часа после их публикации.

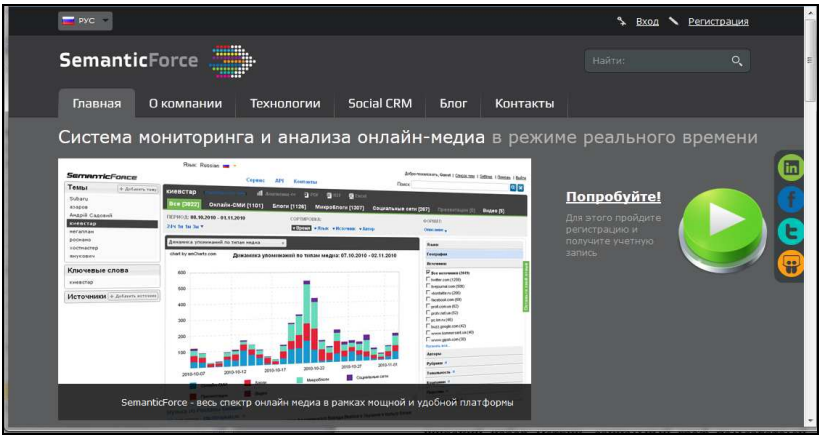


Рис. 12 – Фрагмент веб-сайта службы SemanticForce

Wobot (wobot.ru) – сервис, позволяющий прослеживать ретроспективу мнений в социальных сетях. Доступен широкий набор метрик, социальный граф пользователей. Обладает самообучающимся механизмом, позволяющим определять тональность сообщений.

TweetDeck (tweetdeck.com) – бесплатное кроссплатформенное приложение – инструмент для управления и отслеживания сообщений в социальных сетях Twitter, Facebook, MySpace, LinkedIn. Поддерживает многоканальный колоночный интерфейс, всевозможные фильтры, в том числе, по ключевым словам. Аналитика отсутствует.

1.3.5. Ранжирование источников информации

Мощные возможности Интернет порождают проблему ранжирования информационных источников, оптимизации их состава и количества при

использовании в корпоративной информационной системе с целью обеспечения приемлемого качества, удовлетворяющего потребностям пользователей.

Сегодня становится ясно, что как разработка, так и применение качественно новых средств работы с сетевыми ресурсами переходит в разряд приоритетных задач. В частности, без развитых средств наблюдения за сетевыми информационными источниками невозможно обеспечить соответствующую репрезентативность выборок, а эта задача сегодня является одной из самых актуальных при отборе источников для проведения информационно-аналитической работы.

Существует несколько подходов к ранжированию и отбору информационных источников, остановимся на некоторых детально. Эти подходы базируются на принципе обеспечения максимальной полноты информации при минимальном количестве источников, выборе наиболее оригинальных, тематически стабильных, максимально цитируемых источников.

Системы интеграции и мониторинга новостей из открытых веб-сайтов сети Интернет сегодня все чаще становятся основными компонентами информационных служб различного уровня. Ни для кого уже не секрет, что даже самая закрытая новостная информация, передаваемая информационными агентствами, с минимальной временной задержкой становится доступной в Сети. Можно отметить разнообразный диапазон параметров информационных источников как по объемам публикуемой информации, так и по содержанию – от сообщений информационных агентств – до «живых журналов» (блогов) и социальных сетей.

В этой связи актуальными оказываются вопросы ранжирования и выбора источников новостной информации – веб-сайтов, к которым требуется обеспечить доступ через один интерфейс как в поисковом режиме, так и в режимах аналитического обобщения.

Принципам ранжирования как отдельных веб-документов, так и документальных массивов посвящено большое количество научных работ и практических разработок. Ссылочное ранжирование веб-сайтов сегодня является отдельным направлением интернет-бизеса – SEO (search engine optimization). Вместе с тем, вопросам ранжирования и отбора информационных ресурсов с учетом их новостного контента, объемов и стабильности тематики публикаций уделяется значительно меньшее внимание.

Безусловно, основным критерием при отборе источников для таких систем мониторинга новостей является их содержание. Как уже было замечено, распределение источников по контенту, соответствующему тематическим потребностям корпоративного пользователя удовлетворяет закону Бредфорда, соответственно, при отборе источников обязательно должно учитываться их ранжирование по степени соответствия тематике. Однако реализация такого отбора приводит к известным сложностям. На практике такое ранжирование осуществляется экспертами путем оценивания количества документов, релевантных некоторому отлаженному пакету тематических запросов, адресуемых к фрагменту базы данных, составленной из документов анализируемого источника. А это неизбежно приводит к элементу субъективизма со всеми вытекающими последствиями.

Поэтому представляется перспективным дополнить традиционный подход более объективными и строгими методами, позволяющими оптимизировать процесс формирования информационной базы систем интеграции контента.

Распределение источников по количеству генерируемых документов

В 2008 г. было проведено исследование распределения источников – открытых веб-сайтов – по количеству генерируемых ими документов [Ландэ, 2008-1]. В качестве экспериментального информационного корпуса использовался массив, охватываемый системой

контент-мониторинга InfoStream. В частности, были изучены распределения, относящиеся к массиву документов за март 2008 года объемом свыше 1.2 млн. документов из более чем 2500 источников. На рис. 13 приведен график распределения (в полулогарифмическом масштабе) количества документов, опубликованных источниками, ранжированными по этому параметру. Центральная часть графика хорошо аппроксимируется прямой, что свидетельствует о близости представленной зависимости к гиперболической (т.е. о действии обобщенного закона Ципфа). На рис. 14 приведено общее количество документов, охватываемых системой мониторинга в зависимости от учитываемых в ней источников, также ранжированных по количеству опубликованных документов.

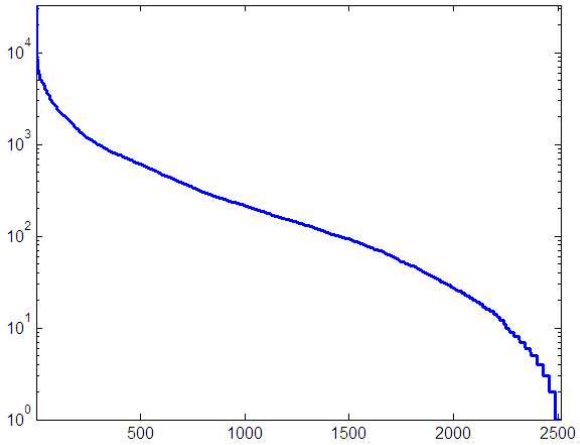


Рис. 13 – Количество публикаций (ось ординат), соответствующих ранжированному списоку источников (ось абсцисс)

Поскольку закон Ципфа предполагает аппроксимацию плотности распределения гиперболической зависимостью вида a/x, то функция распределения количества документов:

$$f(x) \sim \int \frac{a}{x} dx = a \ln x + C$$

в разумном приближении описывается логарифмическим законом.

Приведенная зависимость позволила построить критерий отбора необходимой части источников для различных корпоративных применений из общего списка источников охватываемых системой мониторинга InfoStream, удовлетворительно решающих задачи пользователей.

Если предположить, что все источники давали бы одинаковый вклад по количеству опубликованных документов, то рассматриваемая зависимость была бы линейной и выражалась формулой:

$$f_{lin}(n) = n \frac{f_{\max}}{N},$$

где $f_{\max}$ – максимальный объем охватываемых документов, N - количество источников.

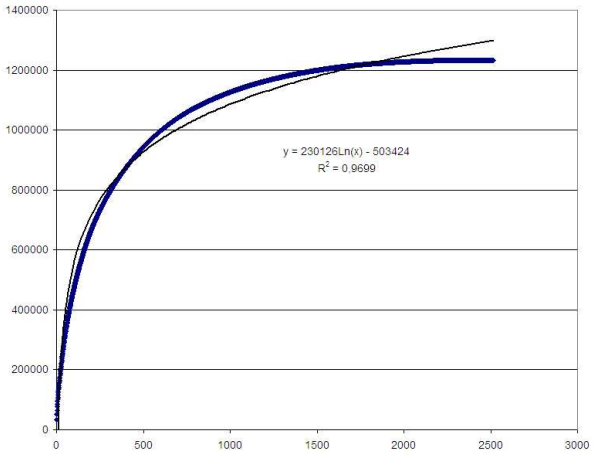


Рис. 14 – Количество публикаций в системе мониторинга (ось ординат) в зависимости от

источников (ось абсцисс), ранжированных по количеству документов

Очевидно, что отклонение реальной зависимости от линейной сначала возрастает, а затем уменьшается до нуля. Будем называть количество источников пороговым n_p, когда значение реальной зависимости $f(n)$ максимально отклоняется от приведенной линейной:

$$n_p = \arg\max \{ f(n) - f_{lin}(n) \}.$$

На рис. 15 приведена иллюстрация значений n_p для различных значений N, т.е. когда выбирается N наиболее продуктивных источников.

Значения n_p практически линейно зависят от N (что вполне соответствует характеру функции $f(n)$, рис. 14): $n_p \sim 0.24N$, при этом количество охватываемых документов, соответствующих n_p при максимальном количестве источников (2514, рис. 15) достигает 80 процентов от $f_{\max}$.

При этом можно заметить, что построенная зависимость удовлетворяет принципу Парето: приблизительно 20% наиболее продуктивных источников публикуют 80% документов.

Наиболее цитируемые источники

Как уже было отмечено, цитируемость отдельных документов и веб-сайтов сегодня является одним из основных критериев оценки рангов документов в сетевых поисковых системах (PageRank, HITS, Salsa, TrustRank, h-индекс и др.) Идея оценки уровня цитируемости позволила построить одну из первых моделей веб-пространства [Broder, 1999]. Главное ее достоинство состоит в том, что она естественным образом сочетает в себе содержательный аспект с возможностью использования количественных

параметров, значения которых определяются вполне объективно. Ниже будет приведена модель новостного информационного пространства, также учитывающая структуру цитирования источников.

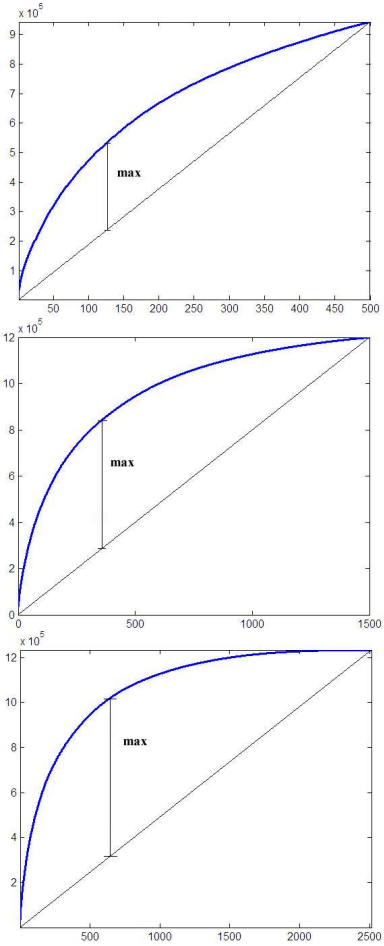


Рис. 15 – Количество публикаций в системе мониторинга при подключении новых наиболее интенсивных источников (500, 1500, 2500)

Выбор наиболее оригинальных источников

Специальное место в технологиях ранжирования источников занимает изучение смыслового дублирования информации. Одной из главных особенностей новостной информации является наличие большого количества сообщений, дублирующих друг друга. Так, о событии мирового значения напишут все СМИ, причем, скорее всего, на одной из первых страниц. Потребитель же (за исключением некоторых специфических направлений аналитических исследований информационного пространства) желает получать по каждому событию одно основное сообщение. Поэтому исследование характера и свойств дублирования информации приобретает в современных технологиях исключительно важное значение.

Тематическая стабильность

Одной из важных характеристик информационных источников в новостном сегменте Сети является их стабильность, понимаемая как генерация постоянного числа документов в единицу времени (естественно, с учетом периодичности изданий). Примером стабильных источников могут служить крупные информационные агентства, регулярно поставляющие потребителям примерно одинаковые объемы информации на протяжении длительного времени, а примером нестабильных – блоги, многие из которых активно действуют в течение нескольких дней, а затем угасают.

Естественно, источник, регулярно выпускающий свою продукцию, с большей вероятностью отразит в своих публикациях важные события, чем источник, выходящий нерегулярно, от случая к случаю (он может попросту «проскочить мимо события»).

С другой стороны, крупные издания, обеспечивающие полноценное освещение нашей жизни, как правило на первое место выводят масштабные, значительные в общественном понимании события, о которых мы все равно так или иначе узнаем, если не из

телевизора, так из разговоров в метро. События же меньшего, так сказать, общественного веса, но при этом, возможно, интересные и важные для отдельных групп потребителей, либо вообще отсутствуют, либо теряются «на последних страницах». Поэтому задача оптимального учета стабильности источников отнюдь не тривиальна и требует, на наш взгляд, серьезных исследований.

Обратимся к одному из важных ее аспектов. Как было показано, ежедневное общее количество документов, публикуемых на основных информационных веб-сайтах приблизительно постоянно и колеблется в основном в зависимости от дня недели. Вместе с тем тематика публикаций подвержена существенным колебаниям.

Один из возможных подходов к решению проблемы ранжирования источников информации основывается на подходе, заключающемся в изучении динамики порождаемых ими тематических информационных потоков.

На практике среди множества проблем подбора и анализа источников контента большое значение, в частности, имеет учет параметров их тематической стабильности. При этом тематическая стабильность и стабильность публикации информации источниками зачастую играют решающую роль при проведении аналитических исследований. Например, такие важные свойства информационных источников, как тематическую корреляцию и полноту, имеет смысл учитывать только для источников, публикующих документы относительно стабильной тематической направленности.

Тематическую стабильность источника можно определить как корреляцию наборов тематических рубрик, которым соответствуют документы из этого источника в различные периоды времени. Можно предположить, что конкретный набор рубрик должен мало влиять на метод расчета стабильности источников.

При исследовании тематической направленности некоторых источников информации могут быть обнаружены документы, отклоняющиеся от основной направленности этих источников. Такие документы, если их количество относительно невелико, не должны влиять на рассчитываемый уровень стабильности источников, в том числе и некоторыми погрешностями в рубрикации при статистическом исследовании можно пренебречь.

Для вычисления уровня стабильности источника информации использовалась формула, основанная на так называемом R/S -анализе [Федер, 1991]. R/S - анализ позволяет исследовать «изрезанность» кривой, образуемой временным рядом на основе отношения разброса значений к среднеквадратичному отклонению.

В [Ландэ, 2008] был предложен параметр тематической стабильности K временного ряда интенсивности публикаций на веб-сайтах (источниках), который выглядит следующим образом:

$$K = \frac{1}{N}\sum_{i=1}^{N}\frac{S_i}{R_i},$$

где N - количество тем (рубрик) источника; S_i – среднеквадратичное отклонение по рубрике i; R_i – размах значений по рубрике i.

Значение S_i вычисляется по формуле:

$$S_i = \sqrt{\frac{1}{M}\sum_{j=1}^{M}\left\{ r_j^{(i)} - \frac{1}{M}\sum_{k=1}^{M} r_k^{(i)} \right\}^2},$$

где $r_j^{(i)}$ – количество вхождения рубрики i за день j, M - количество значений ряда измерения (недель, например).

Значение R_i вычисляется следующим образом:

$$R_i = \max_{1<k<M} X_k^{(i)} - \min_{1<k<M} X_k^{(i)},$$

где $X_k^{(i)}$ – накопленное к моменту k отклонение по рубрике i, вычисляемое по формуле:

$$X_k^{(i)} = \sum_{j=1}^{k} (r_j^{(i)} - \tfrac{1}{M} \sum_{l=1}^{M} r_l^{(i)}).$$

На рис. 16. представлена типовая кривая значений коэффициентов стабильности для источников (было измерено поведение свыше 2500 источников за 2007 год), ранжированных по этим значениям.

Кроме приведенной тематической, может учитываться и более простая диаграмма внетематического распределения источников, ранжированная по коэффициентам стабильности.

Полученные данные еще раз подтвердили тот факт, что электронные издания более склонны изменять тематику публикаций, чем свои объемы, выраженные общим количеством публикаций.

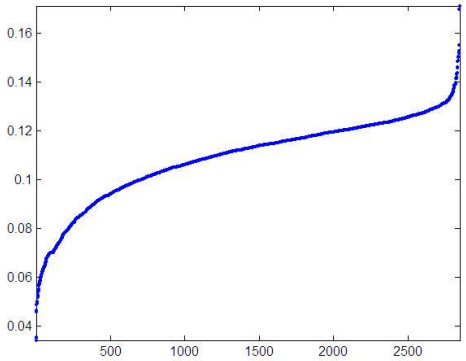


Рис. 16 – Ранжированный список источников (ось 0X) по параметру тематической стабильности (ось 0Y)

Отметим лишь несколько практических применений ранжирования информационных источников. Во-первых, это даст возможность

выявления первоисточников информации, например, для размещения в них рекламных материалов, материалов информационного влияния и т.п. Во-вторых, можно сократить затраты времени и средств путем игнорирования, исключения из поиска и анализа заведомо слабых, «шумовых» источников. Кроме того, для оперативного нахождения актуальной информации корректное ранжирование может способствовать нахождению действительно полезных первоисточников и служб интеграции информации.

Результаты исследований источников информации могут использоваться при ранжировании выдачи информационно-поисковых систем, подсчете медиа-рейтингов, позволяют рекомендовать пользователям наиболее тематически стабильные и оригинальные источники информации, например, для включения их в список «персональных» в интерфейсах систем контент-мониторинга информационных ресурсов.

1.4. Новостной сегмент веб-пространства

1.4.1. Свойства информационного пространства

Отметим основные свойства информационного пространства [Манойло, 2003]:

1. Информационное пространство динамично. В нем не бывает завершенного состояния. Физические объекты, как правило, имеют строго определенные физические пределы. Отсюда возможно следующее следствие: достаточно трудно достичь постоянного информационного доминирования, хотя возможно достижение временного информационного превосходства.
2. Информационное пространство структурировано. Оно неоднородно, в нем есть аттракторы, привлекающие внимание, и барьеры, отталкивающие внимание потребителя

от данной точки информационного пространства.

3. Информационное пространство всегда защищено, в нем есть места, сознательно защищаемые от чужого вхождения. Защита одновременно предполагает наличие слабых мест, служит их детектором.

4. Информационное пространство универсально: любая область человеческой деятельности опирается на него. Отсюда и возникают уникальные возможности для воздействия в любой профессиональной области.

5. Информационное пространство не связано напрямую с реальным пространством из-за его частично нематериальной природы, а также возможности использовать гражданские информационные инфраструктуры, которые достигают любой точки земного шара, тогда как привычные военные методы требуют своих собственных средств.

6. Информационное пространство обладает национально-специфичными способами построения, обработки и распространения информации.

Отсюда вытекают следующие особенности информационного пространства [Манойло, 2003]:

- Информационное пространство структурировано, неоднородно, динамично. В нем есть аттракторы, привлекающие внимание, и барьеры.

- Информационное пространство универсально: любая область человеческой деятельности опирается на него.

- Для информационного пространства характерно четкое разделение таких понятий, как «информация» и «знание», при этом информация

начинает рассматриваться как сырье для производства знаний.

- Информационное пространство является базовым для коммуникаций, формирования, обработки информации и генерации и хранения знаний.
- Информационное пространство – одно из базовых понятий информационных войн, информационных операций. При этом информационную войну можно определять как несанкционированную деятельность в чужом информационном пространстве.
- В информационном пространстве действуют динамические информационные процессы генерации, приема и передачи информации.
- Информационное пространство всегда в какой-то мере защищено, в нем всегда есть места, защищаемые от чужого влияния. Защита предполагает одновременно и наличие слабых мест, служит их детектором.
- Информационное пространство имеет уникальные возможности для воздействия на профессиональную область.
- Информационное пространство технологически основывается на информационных системам организаций, локальных и глобальных сетевых ресурсах.
- Информационное пространство не связано напрямую с реальным пространством из-за его частично нематериальной природы, оно обладает специфичными способами построения, обработки и распространения информации.
- Информационное пространство, как любая сущность имеет жизненный цикл, включающий фазы формирования, развития, рецессии, депрессии.

- Информационное пространство обладает специфичными способами построения, обработки и распространения информации.

Основными структурными составляющими ИП в его синергетическом представлении являются информационные поля и информационные потоки.

Информационное поле – это совокупность всей сосредоточенной в данном объеме пространства-времени информации, безотносительно к ее форме и состоянию, находящейся в отрыве как от объекта отражения, так и от субъекта восприятия. Информационное поле образуется объективной, генетической и идеализированной информацией. Движение информации в информационном поле осуществляется посредством физической связи между реципиентом и источником информации, материализованной в информационном потоке.

Информационный поток – это совокупность информации, перемещающейся в информационном пространстве по каналу коммуникации. Информационные потоки могут протекать как внутри отдельных инфосфер, так и между ними, в зависимости от наличия каналов коммуникации. При этом содержательный характер информационного потока находится в зависимости от характеристик канала коммуникации, так, для передачи информационного потока о графическом объекте необходимо использовать канал коммуникации, обеспечивающий передачу зрительных образов (изображений), в противном случае неизбежны неточности и искажения содержания передаваемой в информационном потоке информации и ее восприятия реципиентом.

Для информационного пространства общества характерны некоторые уникальные субъекты и сообщества, не имеющие прямых аналогов в иных пространствах. К ним относятся:

- социальное виртуальное сообщество;
- онлайновое сообщество;

- сетевой социум;
- виртуальная коалиция.

Одна из важнейших характеристик ИП — его структурированность, под которой понимается такое его свойство, при котором все содержание и особенности этого пространства представляются «информационными компонентами» и взаимосвязями между ними, выраженными в понятном виде. Иными словами, выделены его элементы, установлены связи между ними, введены обозначения, элементы и связи упорядочены.

Виды информационного пространства

Различают пять степеней структурированности информационного пространства: неструктурированные (например, разговорная речь); слабо структурированные (например, письменность); структурированные (например, ИС); формализованные (для которых известны не только информационные компоненты и связи между ними, но и алгоритмы получения значений любого компонента, например технико-экономические показатели деятельности объекта); машиноструктурированные, для которых известны алгоритмы получения не только информационных компонентов, но и их структурных единиц. Информационные компоненты объектов могут иметь различную природу – это документация (организационно-распределительная, экономическая, конструкторская и т. п.), отчеты о НИР, информация на машинных носителях, звуковая и видеоинформация, информация от датчиков и т. д.

Основные функции информационного пространства

Информационное пространство реализует такие основные функции:

1. *Интегрирующая.* В рамках данной функции ИП объединяет в единую пространственно-коммуникативную и социокультурную среду различные

виды человеческой деятельности и занимающихся ими субъектов, в том числе, как отдельных людей, так и целые государства, народы и международные коалиции и транснациональные корпорации.

2. *Коммуникативная.* Информационное пространство создает особую среду трансграничной, интерактивной и мобильной коммуникации различных субъектов деятельности, в рамках которой они осуществляют информационный обмен.

3. *Актуализирующая.* Именно в информационном пространстве осуществляется актуализация интересов различных субъектов деятельности посредством реализации ими информационной политики.

4. *Геополитическая.* Информационное пространство формирует собственные ресурсы и изменяет значимость традиционных ресурсов, создавая новую среду геополитических отношений и конкуренции.

5. *Социальная.* Информационное пространство трансформирует состав общества и изменяет характер и содержание социально-политических (общественных) отношений во всех сферах – политике, культуре, науке, религии и других.

1.4.2. Модель новостного сегмента веб-пространства

Эффективный анализ новостных информационных потоков в сети Интернет, построение систем агрегации новостей невозможны без некоторых сведений о структуре новостного веб-пространства. Это пространство формируется динамичными потоками сообщений, публикуемых на веб-сайтах средств массовой информации, информационных агентств, отдельных организаций, в социальных сетях, блогах. Чем, например, может быть полезно знание о структуре новостного веб-пространства на практике? Во-первых, это даст возможность выявления первоисточников информации, например, для их тщательного анализа или размещения в них рекламных материалов информационного влияния и т.п. Во-вторых, можно

сократить затраты времени и других ресурсов путем игнорирования, исключения из поиска и анализа заведомо слабых, «мусорных» источников. Кроме того, для оперативного нахождения актуальной информации корректная модель может способствовать нахождению действительно полезных первоисточников и служб интеграции информации.

Если для обычного веб-пространства уже признана модель «галстука-бабочки», представленная в работах А. Бредера и его коллег [Broder, 2000] , то архитектура новостного веб-пространства менее изучена. Можно было бы попросту применить модель А. Бредера к новостной составляющей веб-пространства, однако такой подход нельзя считать корректным по ряду причин:

- новостные потоки характеризуются динамикой, что сильно влияет на природу гиперссылок. Например, на наиболее актуальные сообщения в течение определенного времени ссылок может вообще не существовать;
- модель Бредера слабо учитывает особенности глубинного веб, т.е. тех информационных веб-ресурсов, на которые не существует прямых гиперссылок (в рассмотрение им брались ресурсы, уже охваченные поисковой системой AltaVista);
- в новостных потоках необходимо учитывать не только гиперссылки, но и ссылки контекстные, причем не только на объекты из открытой части веб-пространства (это могут быть зачастую ссылки на ресурсы, доступные только по паролю, или даже оффлайновые публикации изданий, возможно и присутствующих в Интернет);
- модель Бредера не включает такого понятия, как содержательное дублирование информации;
- при построении модели структуры новостного веб-пространства наибольшее внимание должно

уделяться именно веб-ресурсам (веб-сайтам), на которых публикуются новостные сообщения, а не отдельным веб-страницам, публикациям в блогах или самим сообщениям.

В работе [Ландэ, 2006] была построена модель новостного пространства, в качестве экспериментальной базы для этой модели использовалась ретроспективная база данных системы контент-мониторинга InfoStream [Григорьев, 2005] на то время объемом 30 млн. документов из 2000 источников.

В результате обработки специального пакета запросов к ретроспективной базе данных для каждого сообщения, относящегося к определенному источнику информации, были выявлены исходящие ссылки на другие источники (ссылки на собственный источник исключались).

Специальное место в исследовании занимало изучение смыслового дублирования информации. При этом следует отметить, что процент дублирующихся сообщений в системе InfoStream значительно меньше, чем во всем веб-пространстве. Это объясняется подбором источников для сканирования, в число которых не входят многие новостные интеграторы.

Выявление дублирующихся по содержанию новостных сообщений в системе InfoStream выполняется на основе лингвостатистических методов, заключающихся в выявлении и последующем сравнении наиболее весомых слов в документах, которые выступают своеобразными ключами [Ландэ, 2007], [Ландэ, 2014].

Было проведено исследование соотношения дублирующихся и оригинальных сообщений, которые привели к неожиданному результату. Оказалось, что количество оригинальных сообщений и их содержательных дублей, охватываемых системой InfoStream, почти в точности совпало (рис. 17).

Следует заметить, что устранение дублирующихся

сообщений в информационных потоках требуется далеко не всегда. Существует ряд задач, в которых используется факт дублирования текстов сообщений в различных источниках, например при определении важности сообщения (если сообщения многократно дублируется на сайтах и в СМИ) или при определении эффективности PR-кампаний (подсчет републикаций пресс-релизов и др.)

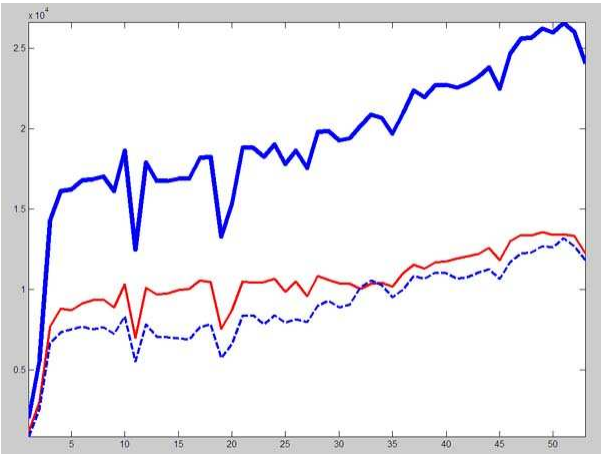


Рис. 17 – Объемы информации, сканируемой системой InfoStream в 2005 г., в разрезе недель: жирная линия – общий объем, тонкая – оригинальные сообщения, пунктирная – информационные дубли

В результате проведенных исследований была предложена модель новостного веб-пространства, которая представлена на рис. 18. Эта модель включает такие зоны:

- *входной полуостров*. Веб-сайты, которым соответствуют менее порогового значения входящих ссылок и любое превышающее пороговое количество исходящих ссылок (16,7% веб-сайтов);
- *выходной полуостров*. Веб-сайты, которым соответствуют менее порогового значения исходящих ссылок и любое превышающее

пороговое количество входящих ссылок (27,5% веб-сайтов);

- *остров*. Веб-сайты, которым соответствуют менее порогового значения исходящих и входящих ссылок (19,3% веб-сайтов);
- *ядро*, состоящее из трех областей: входной, выходной и коммуникационной зоны (36,5% веб-сайтов). Зона ядра характеризуется средними и большими значениями уровней исходящих и входящих связей, однако, как видим, допускает ранжирование по уровню этих коммуникаций.

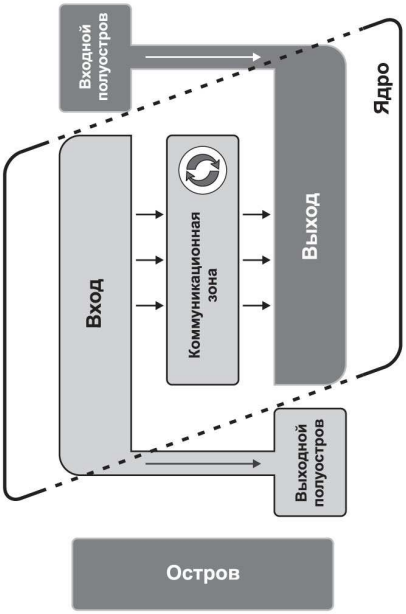


Рис. 18. Архитектура новостного веб

Основа модели была построена путем анализа полной картины распределения входных и выходных ссылок. При этом строилась матрица инциденций и соответствующие графы связи, а также выявлялись необходимые кластеры. Вместе с тем оказалось, что

само по себе отношение количества входящих и исходящих ссылок для каждого из источников достаточно точно характеризует его попадание в названные кластеры.

Проведенные исследования и полученные закономерности послужили основой «идеальной схемы» новостного веб-пространства (рис. 19), основанной на контекстных ссылках.

Рис. 19 – Представление областей модели в зависимости от количества исходящих и входящих ссылок

Вместе с тем данная модель предполагает дальнейшее развитие в следующих направлениях: более точной идентификации контекстных ссылок, совершенствования критерия определения зон на основе полного учета структуры ссылок и методов кластерного анализа, совершенствования механизма определения содержательного дублирования информации (в том числе за счет механизмов настройки сканеров системы контент-мониторинга, учета авторитетности источников и возможных умышленных задержек публикации в Интернет).

1.5. Модели информационных сюжетов

Определим информационный сюжет (далее – инфосюжет) как множество документов (информационных сообщений), посвященных одной тематике или одному событию. «Средой обитания» инфосюжетов является информационное пространство, сегодня вполне репрезентативно представляемое сетью Интернет, что однако не ограничивает авторов рассмотрением только этой сети. Инфосюжеты можно трактовать как документальные или контентные системы (от англ. content – содержание), которые, как будет показано ниже, полностью удовлетворяют общему определению систем.

Одним из важнейших свойств информационных сюжетов является их живучесть. Очевидно, что живучесть инфосюжета может, с одной стороны, рассматриваться как его объективное свойство, которое зависит от тематики, аудитории, времени, а с другой стороны, как характеристика, которую хотят придать ему в случае искусственного формирования, например, в ходе проведения информационных операций.

Известно, что системы в общем случае могут разделяться на целенаправленные и нецеленаправленные. Инфосюжеты как системы могут относится как к первому, так и ко второму классу. Живучесть искусственно формируемых инфосюжетов имеет решающее значение, например, в ходе рекламных кампаний и других информационных операций.

Обратимся к традиционному пониманию понятия живучести. Живучесть – это фундаментальное свойство сложных систем. Биологические, социальные и многие другие системы изначально обладают свойством живучести, что позволяет им сохранять целостность, выполнять свои функции и развиваться вопреки деградации и независимо от наличия неблагоприятных воздействий со стороны внешней среды [Shelton, 2003]. Методы и средства обеспечения живучести успешно применяются при создании сложных организационно-

технических систем.

Инфосюжеты конечно же нельзя считать ни биологической, ни технической системой, хотя отдельные элементы этих систем необходимы для их существования. Скорее всего инфосюжеты можно отнести к системам коммуникационным, на формирование которых существенное влияние оказывает так называемый «человеческий фактор», который, пожалуй, сложнее всего поддается формализации.

Живучие системы способны поддерживать непрерывное выполнение своих основных функций (в случае инфосюжетов – информировать о наиболее важных аспектах тематики или события), временно или постоянно отказываясь от выполнения менее важных функций информирования, изменять свою структуру и поведение, находить и выполнять новые функции, необходимые для успешного противостояния неблагоприятным воздействиям из внешней среды (информационного пространства), приспосабливаясь к условиям своего функционирования [Додонов, 1990].

Для того, чтобы перейти к рассмотрению инфосюжета как системы и сопоставить его свойства со свойствами остальных систем, обратимся к классическому определению, в соответствии с которым система – это совокупность объектов и связей между ними, выделенных из среды на определенное время и с определенной целью. Система в общем смысле рассматривается как динамически изменяемая совокупность сильно связанных объектов, обладающая свойствами организации, связности, целостности и членимости (декомпозиции).

Соответственно, информационный сюжет можно трактовать как контентную систему, совокупность документов, связанных взаимными контекстными ссылками, гиперссылками, цитированием, общей лексикой, фактографией и т.д., выделенных из информационной среды (информационного пространства) на определенное время (время

актуальности) с определенной целью или по определенному поводу. Последнее относится к событиям, которые можно трактовать как нецеленаправленные системы. То есть действительно, инфосюжет – это совокупность сильно связанных объектов, обладающая свойствами организации, связности, целостности (определяемой тематикой или событием) и членимости (на отдельные документы, их аспекты).

Выделяют такие свойства систем, связанные с их целями и функциями:

1. Синергетичность – однонаправленность действий компонентов усиливает эффективность функционирования системы. В случае инфосюжетов направленность отдельных документов усиливает информационную функцию всего инфосюжета.

2. Приоритет интересов системы перед интересами ее компонент (общую тематическую тенденцию определяет весь инфосюжет, а не отдельные документы как компоненты).

3. Эмерджентность – цели (информационные функции) компонент (отдельных документов) системы не всегда совпадают с целями (функциями) всего инфосюжета.

4. Мультипликативность – и позитивные, и негативные эффекты функционирования компонент системы обладают свойством умножения, а не сложения (аналогии – количество информации в документах, информационная энтропия).

5. Целенаправленность инфосюжетов в случае их искусственного формирования (вместе с тем, существуют и нецеленаправленные системы, в том числе и инфосюжеты).

6. Альтернативность путей функционирования и развития. Важнейшие документы могут быть

актуальными на протяжении длительного времени или отдельные документы по одной тематике, генерируемые в большом количестве, но имеющие небольшой срок актуальности.

Информационные сюжеты обладают следующими свойствами, связанными с их структурой:

1. Целостность – первичность целого инфосюжета по отношению к отдельным документам.
2. Неаддитивность – принципиальная несводимость свойств инфосюжетов к сумме свойств составляющих их документов.
3. Структурность – возможна декомпозиция инфосюжета на компоненты (документы), установление связей между ними.
4. Иерархичность – компоненты системы (информационные сообщения, документы, пожалуй, кроме элементарных одноаспектных), могут также рассматриваться как подсистемы инфосюжета.

Инфосюжеты, как и отдельные документы, являются элементами информационного пространства, и им, соответственно, присущи свойства, связанные с взаимодействием с внешней средой:

1. Коммуникативность – существование сложной системы коммуникаций инфосюжетов с информационным пространством, в частности, отдельные документы из инфосюжета могут быть связаны не только с другими документами из того же инфосюжета, но и с другими частями информационного пространства.
2. Взаимодействие и взаимозависимость инфосюжета и информационного пространства.
3. Адаптивность – стремление к состоянию устойчивого равновесия, которое предполагает адаптацию параметров инфосюжета на

определенных этапах его жизненного цикла к изменяющимся параметрам внешней среды.

4. Надежность – существование инфосюжета при выходе из строя отдельных его компонент (документов), сохраняемость значений параметров системы в течение определенного периода.
5. Интерактивность – взаимодействие с внешней средой и «ответная» изменяемость инфосюжетов.

Существует еще ряд системных свойств инфосюжетов, среди которых можно назвать:

1. Интегративность – наличие системообразующих, системосохраняющих факторов.
2. Эквифинальность – способность инфосюжетов достигать состояний, не зависящих от исходных условий и определяющихся только параметрами системы.
3. Наследственность.
4. Развитие.
5. Самоорганизация и т.д.

Для полноценной работы или сохранения минимального набора критически важных функций информирования инфосюжет должен обладать вполне определенным запасом устойчивости к дестабилизирующим воздействиям из внешней среды (информационного пространства), обусловленных, в свою очередь, влияниями со стороны общества, государства, коммерческих структур и т.д. Как на весь информационный сюжет, так и на отдельные документы, входящие в него, могут оказываться различные дестабилизирующие информационные воздействия, атаки, например, удаление отдельных материалов с веб-сайтов сети Интернет, уничтожение или отключение информационных серверов, публикация новых документов, которые в определенном направлении исказят исходный инфосюжет или

порождение нового инфосюжета, который может снизить актуальность или попросту уничтожить исходный сюжет.

Понятно, что для полноценной работы и сохранения минимального набора критически важных функций информирования инфосюжет должен обладать вполне определенным запасом устойчивости к внешним дестабилизирующим воздействиям. При этом нарушение целостности инфосюжета на фоне снижения актуальности его компонент влечет за собой дезорганизацию, одновременную потерю гибкости – понижение живучести и нарушение целостности, т.е. потерю важнейших функций инфосюжетов (рис. 20).

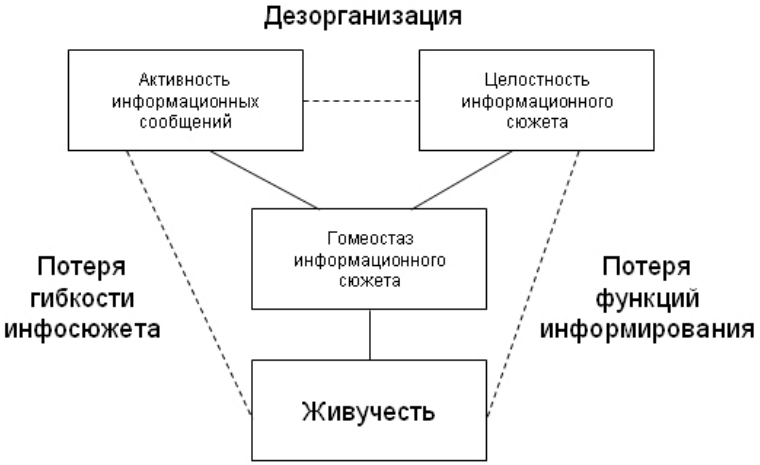


Рис. 20 – Модель гомеостаза инфосюжета

Информационные сюжеты могут быть представлены как сетевые структуры, так называемые динамические сети. Текущее состояние инфосюжета может быть представлено в виде графа $< M, L >$, где M – это множество документов, входящих в сюжет, а L – множество ребер – связей подобия, цитирования, ссылок и т.д. Свойство живучести напрямую связано с такими свойствами графов, как связность,

кластерность, средний кратчайший путь между вершинами и т.п.

В инфосюжетах важно ранжировать информационные сообщения, выделять основные из них. Тут на помощь могут прийти такие популярные методы ранжирования, как PageRank, HITS, Salsa, а также параметры, вычисляемые в рамках теории сложных сетей (complex networks).

К потере живучести инфосюжета может привести разрыв связей между его компонентами, например, устраненим из информационного пространства наиболее весомых компонент, то есть таких, которые имеют наибольший коэффициент посредничества (betweenness). Этот коэффициент для конкретного узла сети определяется как сумма по всем парам узлов сети соотношений количества кратчайших путей между ними, проходящими через заданный узел, к общему количеству кратчайших путей между ними.

Как и сеть террористов, восстановление которой после деструктивного воздействия описано в [5], информационный сюжет также является динамической системой, восстановление которой после уничтожения лучших «посредников» осуществляется за счет латентных связей с другими компонентами информационного пространства. После того, как инфосюжет разделяется на изолированные фрагменты, он может использовать эти связи и быстро восстановить связность (рис. 21).

Воссоединение частей сети не состоится, если ни одна из пар уцелевших после деструктивного воздействия компонент не сможет найти скрытые связи между собой (возможно, не прямые, а через другие компоненты информационного пространства). В этом случае влияние разъединения на показатели функционирования инфосюжета зависит от того, смогут ли снова разъединенные части получить взаимные связи, недостаток которых наблюдается в этой части инфосюжета.

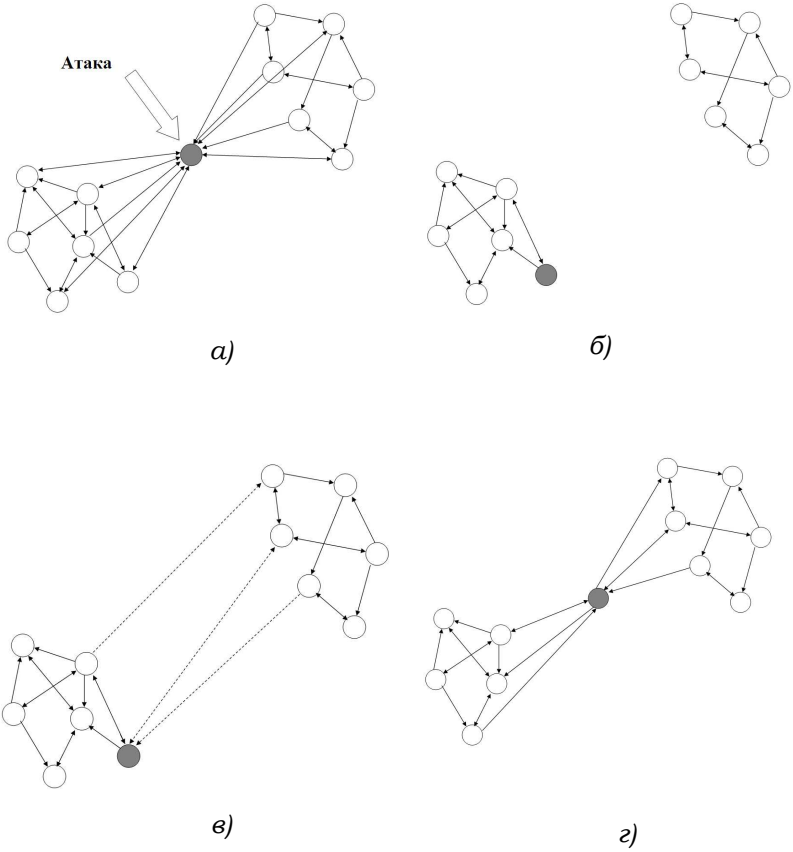


Рис. 21 – Восстановление структуры сети путем выбора нового «посредника»: а) - атака на сеть; б) - несвязная сеть с изъятым посредником; в) у - возрождение скрытых (латентных) связей; г) - связность сети восстановлена

Если часть инфосюжета была близкой к самодостаточности, то она продолжает функционировать самостоятельно. В противном случае, она прекращает функционирование до тех пор, пока не сформируется новая связь. Если одно из соединений оказывается успешным, то его инициатор становится

новым «посредником», который объединяет две части сети.

Воссоединение частей сети не состоится, если ни одна из пар уцелевших после деструктивного воздействия компонент не сможет найти скрытые связи между собой (возможно, не прямые, а через другие компоненты информационного пространства). В этом случае влияние разъединения на показатели функционирования инфосюжета зависит от того, смогут ли снова разъединенные части получить взаимные связи, недостаток которых наблюдается в этой части инфосюжета.

На практике при поиске новостной информации всегда возникает задача выявления инфосюжетов, состоящих из отдельных документов, и их ранжирования по некоторым признакам, что должно обеспечить не только выявление самой важной темы, но и "веерное" многоаспектное освещение всех наиболее значимых аспектов инфосюжетов. Эта задача решается во многих системах с использованием различных подходов и алгоритмов. При этом неизменной остается технологическая цепочка: построение семантической сети из информационных сообщений, кластеризация – выявление наиболее взаимосвязанных групп, то есть инфосюжетов, «взвешивание» (оценка важности, актуальности) и наглядная визуализация самых весомых из них.

При выделении сюжетных цепочек для определения попарной текстуальной близости отдельных документов, как правило, используются алгоритмы выявления подобных документов, ставшие уже традиционными в поисковых системах. Так матрица попарной близости документов обрабатывается алгоритмами кластеризации, такими как *LSA/LSI*, *k-means*, методом суффиксных деревьев и т.д. Выделенные классы документов и представляют собой инфосюжеты.

Для предъявления пользователям инфосюжеты должны быть ранжированы. Основные факторы,

влияющие на ранжирование по важности - оперативность информации и размер сюжетной цепочки. Под оперативностью информации понимается некоторая функция от времени публикации всех документов в инфосюжете, а размер отражает общий интерес к конкретной теме. Во всех этих подходах центральная задача состоит в отождествлении документов, относящихся к одному сюжету и выделение «непересекающихся» сюжетов.

Следует обратить внимание на то, что проблема автоматического построения качественных инфосюжетов на основе тематических потоков сетевой новостной информации сегодня практически решена.

Для изучения проблем живучести инфосюжетов как сложных многопараметрических систем, параметры которых еще мало изучены, наиболее подходящей методикой является математическое моделирование. Жизненный цикл информационных сюжетов может описываться, например, моделью диффузии информации [Ландэ, 2007-1]. Напомним, что в естественных науках под диффузией понимают взаимное проникновение друг в друга соприкасающихся веществ, вызванное, например, тепловым движением их частиц. Процессы диффузии информации, как и процессы диффузии в физике, достаточно точно моделируются с помощью методов клеточных автоматов.

Клеточные автоматы являются полезными дискретными моделями для исследования динамических систем. Дискретность модели, а точнее, возможность представить модель в дискретной форме, может считаться важным преимуществом, поскольку открывает широкие возможности использования компьютерных технологий.

Модель диффузии информации, которую будет рассматриватьс ниже, является двумерной, поэтому вся система клеточных автоматов для этого случая будет описываться двумерным массивом. В случае двумерной решетки, элементами которой являются квадраты,

ближайшими соседями, входящими в окрестность элемента $y_{i,j}$, можно считать или только элементы, расположенные вверх-вниз и влево-вправо от него, либо добавленные к ним еще и диагональные элементы (окрестность Мура). В модели Мура каждая клетка имеет восемь соседей. Это позволяет определять общее соотношение значения клетки на шаге $t+1$ по сравнению с шагом t:

$$y_{i,j}(t+1) = F(y_{i-1,j-1}(t), y_{i-1,j}(t), y_{i-1,j+1}(t), y_{i,j-1}(t), y_{i,j}(t), y_{i,j+1}(t), y_{i+1,j-1}(t), y_{i+1,j}(t), y_{i+1,j+1}(t)).$$

В рамках данной модели, которая относится к распространению новостей в информационном пространстве, применяются окрестность Мура и вероятностные правила распространения новостей по заданной тематике. Предполагается, что клетка может быть в одном из трех состояний: 1 – «свежая новость» (клетка окрашивается в черный цвет); 2 – новость, устаревшая, но сохраненная в виде сведений (серая клетка); 3 – клетка не имеет информации, переданной новостным сообщением (клетка белая, информация не дошла или уже забыта).

Модель диффузии информации предполагает следующие правила развития информационного сюжета:

- изначально все поле состоит из белых клеток за исключением одной – черной, которая первой «приняла» новость;
- белая клетка может перекрашиваться только в черный цвет или оставаться белой (она может получать новость или оставаться «в неведении»);
- белая клетка перекрашивается, если выполняется условие: $C \cdot pm > 1$, где p – псевдослучайная величина ($0 < p < 1$), m – количество черных клеток в окрестности, C – константа ($C = 1{,}5$ при $m = 1$; $C = 1$ при $m \neq 1$);

- если клетка черная, а вокруг нее исключительно черные и серые, то она перекрашивается в серый цвет (новость устаревает, но сохраняется как сведения);
- если клетка серая, а вокруг нее исключительно серые и черные, то она перекрашивается в белый цвет (происходит забывание сведений при их общеизвестности).

Описанная система клеточных автоматов вполне реалистично отражает процесс развития инфосюжета (рис. 22).

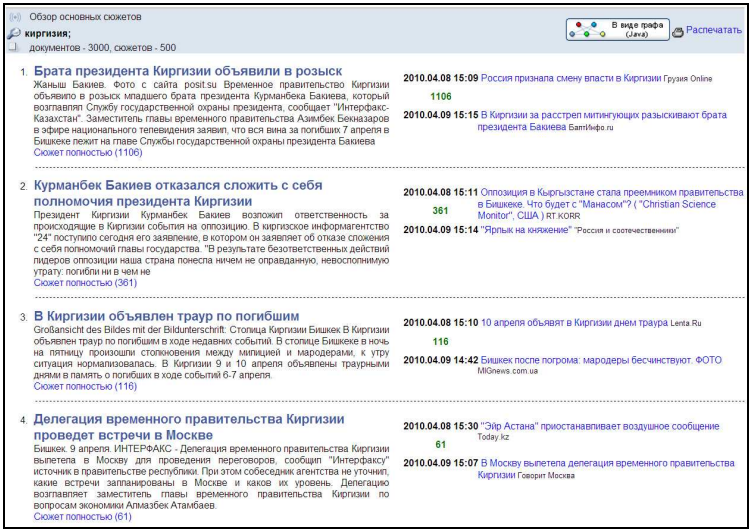


Рис. 22 – Пример отображения инфосюжетов в системе InfoStream

На поле размером 40 х 40 (рис. 23, размеры были выбраны исключительно для наглядности) состояние системы клеточных автоматов полностью стабилизируется за ограниченное количество тактов, т.е. на практике процесс оказался сходящимся.

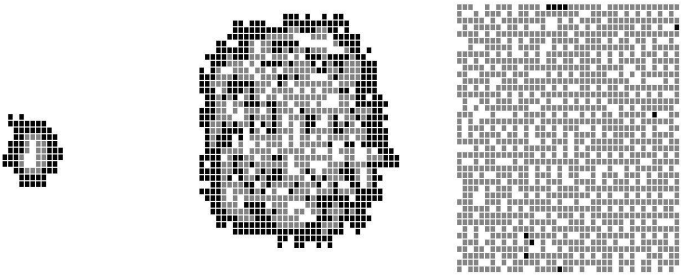


Рис. 23 – Состояния эволюции системы клеточных автоматов

Типичные зависимости количества клеток (последовательности количества однотипных клеток), пребывающих в различных состояниях, в зависимости от шагов итерации приведены на рис. 6.

При анализе приведенных графиков следует обратить внимание на такие особенности: 1 – суммарное количество клеток, пребывающих во всех трех состояниях на каждом шагу итерации постоянно и равно размеру поля; 2 – при стабилизации клеточных автоматов соотношение количества серых, белых и черных клеток приблизительно составляет: 0.75 : 0.25 : 0; существует точка пересечения кривых, определяемых всеми тремя последовательностями на уровне 33 % каждая.

Именно черные клетки образуют актуальный инфосюжет, динамика которого представлена на рис. 24.

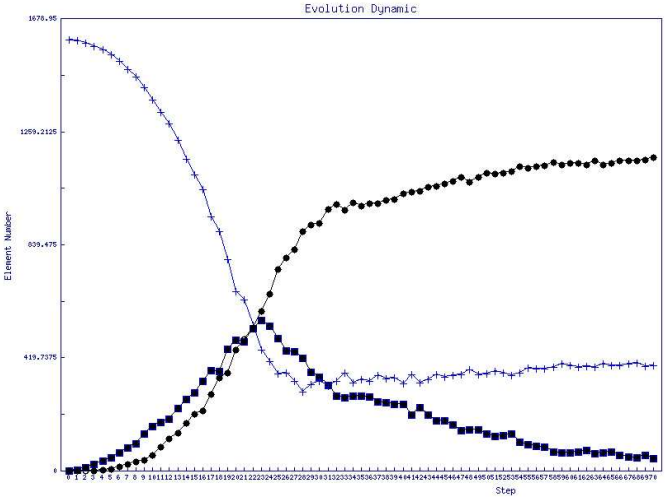


Рис. 24 – Распределение клеток в зависимости от такта системы клеточных автоматов: белые клетки - (+); серые клетки - (•); черные клетки – (■)

Полученные в результате аналитического моделирования зависимости количества серых x_g, белых x_w и черных x_b клеток от шага эволюции системы клеточных автоматов, выражаются формулами:

$$x_g = \frac{0.75}{1+e^{-0.15(t-30)}};$$

$$x_w = 1 - \frac{0.75}{1+e^{-0.25(t-20)}};$$

$$x_b = 0.75\left(\frac{1}{1+e^{-0.25(t-20)}} - \frac{1}{1+e^{-0.15(t-30)}}\right).$$

Жизненный цикл инфосюжетов, также как и многих других систем, может быть описан с помощью еще двух больших классов моделей: булевых и

марковских.

Следует отметить, что зависимость диффузии новостей, полученная в результате моделирования, хорошо согласуется с реальным поведением тематических информационных потоков на интернет-источниках, а на локальных временных промежутках - с традиционными моделями.

В булевой модели можно предположить, что инфосюжет состоит из n элементов (документов), при этом i-му элементу соответствует булева переменная x_i, которая может принимать значения {0, 1}, то есть:

$$x_i = \begin{cases} 1, \text{ если элемент } i \text{ активен;} \\ 0, \text{ иначе.} \end{cases}$$

Принимая во внимание тот факт, что информационное пространство как система динамическое, можно зафиксировать значение n как заведомо большое число, превышающее максимально наблюдаемое количество документов в информационных сюжетах. Несуществующим (недостающим) документам можно присвоить нулевые значения x_i.

Состояние информационного сюжета определяется структурной булевой функцией работоспособности (действенности), зависящей от переменных $x_1, x_2, \ldots, x_n$:

$$S(x_1, x_2, \ldots, x_n) = \begin{cases} 1, \text{ если инфосюжет активен;} \\ 0, \text{ иначе.} \end{cases}$$

Если активность элемента инфосюжета (документа) рассматривать как функцию времени, то состояние i-го документа можно рассматривать как случайный процесс $x_i(t)$, принимающий в произвольные моменты времени $t \geq 0$ значения 0 и 1.

Для инфосюжета как для системы определяется вероятность его работоспособности по известным правилам [Райншке, 1979], [Райншке, 1988].

Среди недостатков булевых моделей можно назвать предположение только о двух состояниях компонентов – активности и неактивности. Кроме того, булевы модели не учитывают то, что весьма существенную роль может играть последовательность, в которой отказывают отдельные компоненты. Кроме того, в общем случае характер отказов отдельных компонент зависит от состояния других компонент. Это находится в противоречии с изначально предполагаемой независимостью элементов в булевой модели.

Информационный сюжет можно описать также марковской моделью. Пусть система имеет m возможных состояний. Обозначим множество состояний через $M = \{z_1,\ z_2,\ \ldots,\ z_m\}$. Для любого фиксированного момента времени $t \geq 0$ состояние системы $z_i(t)$ интерпретируется как случайная величина. Заданы множество всех состояний *M*, вектор распределения начальных вероятностей *p*(0) и функция переходных вероятностей. Определяется вероятность актуальности инфосюжета в заданный момент времени *t* (работоспособность системы).

Применимость марковских моделей также имеет свои границы. Интенсивности переходов между отдельными состояниями системы могут быть нестационарными, принимаемые при расчете допущения относительно распределения интенсивности отказов могут значительно снизить точность полученных результатов; число состояний системы может быть так велико, что расчет становится практически невозможным.

Проведя оценку надежности компонентов системы и получив общие показатели ее надежности, можно оценить ее живучесть на всех этапах их жизненного цикла. Существует несколько подходов, к

проведению оценки живучести, имеющих общий характер. Живучесть системы можно оценить относительно некоторого стандартного внешнего воздействия либо относительно множества внешних воздействий [Додонов, 2004] .

Пусть $E = \{e_i\}$ – множество деструктивных воздействий на инфосюжет;

$\sigma_j(e_i)$ – показатель эффективности (качества) функционирования j-го варианта инфосюжета при воздействии e_i;

$H_j(E) = \min_{e_i \in E} \sigma_j(e_i)$ – показатель живучести инфосюжета для множества возможных воздействий на него *E*.

Тогда при целенаправленном формировании инфосюжета задача проектирования состоит в том, чтобы из множества вариантов инфосюжетов Ω найти такой, для которого выполняется:

$$H_k(E) = \max_{X_j \in \Omega} \min_{e_i \in E} \sigma_j(e_i).$$

Кроме необходимости сохранения множества функций инфосюжетов при неблагоприятных для информационного сюжета воздействях, часто ставится задача сохранения определенного уровня его эффективности (актуальности, информативности).

Для количественной оценки живучести существуют многочисленные подходы, наиболее распространенный среди которых заключается в определении соотношения количества функциональных (работоспособных) состояний системы к общему возможному количеству состояний системы, возникающих при деструктивных воздействиях.

В качестве простого примера рассмотрим инфосюжет, состоящий из четырех документов (n = 4).

Деструктивное воздействие на инфосюжет – устранение из информационного пространства входящих в него документов. Причем первый из документов считается определяющим – его устранение из информационного пространства фактически ведет к потере информационной функциональности всего инфосюжета. Остальные три документа считаются равноправными. Устранение любых двух из них ($k = 4$) также ведет к потере функциональности инфосюжета.

Если обозначить состояние инфосюжета 4-х элементным кортежем, то множество неработоспособных состояний можно представить как объединение двух подмножеств, первое из которых соответствует состояниям с устраненным первым документом, а второе – с актуальным первым документом, но отсутствующими двумя другими.

Мощность первого подмножества составляет $2^{n-1} = 8$, перечислим его компоненты:

(0, 0, 0, 0)

(0, 0, 1, 0)

(0, 0, 0, 1)

(0, 0, 1, 1)

(0, 1, 0, 0)

(0, 1, 1, 0)

(0, 1, 0, 1)

(0, 1, 1, 1)

Мощность второго подмножества составляет $C_n^k + 1 = 4$, его компоненты:

(1, 0, 0, 0)

(1, 0, 1, 0)

(1, 0, 0, 1)

(1, 1, 0, 0)

Мощность всего множества состояний после деструктивного воздействия составляет $2^n - 1 = 15$.

Таким образом, живучесть G инфосюжета составляет:

$G = (15 - 8 - 4) / 15 = 0{,}2$.

Рассмотрен случай, когда все состояния инфосюжета после деструктивного воздействия равнозначны, то есть равновероятны.

Если состояния инфосюжета не являются равновероятными, то живучесть G инфосюжета составляет:

$$G = \frac{\sum_{i=1}^{m} p_i - \sum_{i=1}^{j} p_i^{(0)} - \sum_{i=1}^{l} p_i^{(1)}}{\sum_{i=1}^{m} p_i} = 1 - \sum_{i=1}^{j} p_i^{(0)} - \sum_{i=1}^{l} p_i^{(1)},$$

где m – мощность всего множества состояний после деструктивного воздействия ($m = 2^n - 1$);

j – мощность подмножества состояний с устраненным первым сюжетом;

l – мощность подмножества состояний с актуальным первым сюжетом ($l = C_n^k + 1$);

p_i – вероятность i-го состояния после деструктивного воздействия;

$p_i^{(0)}$ – вероятность i-го состояния после деструктивного воздействия и устранения первого сюжета;

$p_i^{(1)}$ – вероятность i-го состояния после деструктивного воздействия и сохранения первого сюжета, но при потере общей актуальности инфосюжета.

Более общую оценку живучести инфосюжета можно построить исходя из цели его функционирования, множества задач информирования $Q=\{q_1,\ q_2,\ \ldots,\ q_m\}$ и множества компонент (документов). Действительно, любая задача $q_i \in Q,\ \ i=\overline{1,m}$, характеризуется набором элементарных функций (информирования об отдельных аспектах) $F_i=\{f_{j_1}, f_{j_2},\ldots, f_{j_k}\},\ 1\le j_k \le n$, из которых строятся решения этой задачи.

Обозначим через $F=\bigcup_{i=1}^{m} F_i$ множество наборов элементарных функций информирования инфосюжета. Для каждой задачи информирования q_i задается характеристика эффективности решения. Введем функцию потенциальных возможностей функциональных модулей $\varphi:\{1,\ 2,\ \ldots,\ p\}\to P(F)$, где $P(F)$ - множество всех подмножеств F.

Для характеристики возможных конфигураций инфосюжета введем матрицу потенциальных возможностей системы:

$$a_{ij}=\begin{cases}1, \text{ если } f_i\in\varphi(j),\\ 0, \text{ если } f_i\notin\varphi(j),\ j=\overline{1,p},\ i=\overline{1,n}.\end{cases}$$

Текущую конфигурацию инфосюжета будем характеризовать тем, на выполнение каких функций информирования нацелен каждый модуль. Введем двоичную матрицу B размерности $n\times p$ – матрицу текущей конфигурации системы, такую что:

$$b_{ij}=\begin{cases}1, \text{ если модуль } S_j \text{ выполняет функцию } f_i,\\ 0, \text{ в противном случае.}\end{cases}$$

Определим функцию эффективности модулей $\varphi_{эф} : I_s \times I_f \times B \to T$, где $I_s = \{1,\ 2,\ \ldots,\ p\}$ – множество индексов модулей; $I_j = \{1,\ 2,\ \ldots,\ n\}$ – множество индексов элементарных функций; B - множество матриц конфигураций; T – числовое множество количественных мер эффективности (например, размер аудитории, читающей документы инфосюжета и т. п.). Если $\varphi_{эф}(i, j, B) = t_{ij}$, то в конфигурации, определенной матрицей B, модуль S_i выполняет функцию f_j с эффективностью t_{ij}, $\varphi_{эф}(i, j, B) = 0$, если модуль S_i не выполняет функцию f_j.

Для характеристики инфосюжета введем понятие характеристического вектора состояния - n-мерного вектора (n - мощность множества элементарных функций системы). Начальной конфигурации инфосюжета при условии, что выполняется все множество функций F, будет соответствовать характеристический вектор состояния $(0,\ 0,\ \ldots,\ 0)$. Некоторой текущей конфигурации инфосюжета будет соответствовать характеристический вектор $(d_1, d_2, \ldots, d_i, \ldots, d_n)$, где d_i – число «отказов» функции $f_i \in F$. Под «отказом» функции $f_i \in F$ понимается невозможность выполнения функции информирования f_i, то есть d_i – количество реконфигураций инфосюжета из-за «отказа» функции информирования $f_i \in F$.

Решим задачу нахождения множества характеристических векторов состояний инфосюжета, в которых реализуется конфигурация, обеспечивающая выполнение цели функционирования. *Мощность этого множества также может служить мерой живучести*

системы.

Поставленную задачу можно решить в два этапа.

1. Нахождение множества характеристических векторов состояний инфосюжета S_f, определяющих состояния, в которых возможен выбор конфигурации, обеспечивающей выполнение множества элементарных функций *F*. Пусть некоторый начальный инфосюжет характеризуется матрицей B_0. Первоначальную конфигурацию можно построить, исходя, например, из следующих предположений: каждый модуль выполняет только одну функцию, и каждая функция выполняется только одним модулем, то есть:

$$\sum_{j=1}^{p} b_{ij} = 1, \forall i = \overline{1,n},$$

$$\sum_{i=1}^{n} b_{ij} = 1, \forall j = \overline{1,p}.$$

В качестве критерия для задачи оптимизации естественно выбрать:

$$\Phi = \sum_{i=1}^{n}\sum_{j=1}^{p} b_{ij}\varphi_{эф}\left(i, j, B_0\right) \to \min(\max).$$

В зависимости от конкретного смысла функции эффективности справедлива задача нахождения либо максимума критерия Φ, либо его минимума.

Приведенные выше выражения описывают задачу комбинаторного типа, которую можно решить, например, венгерским методом или с помощью эвристического алгоритма.

Предположим, что возникают «отказы» функций, то есть изменяется состояние инфосюжета. Новому состоянию, с учетом имевших место отказов, соответствует характеристический вектор состояния $\left(d_1,\ d_2,\ \ldots, d_i,\ \ldots, d_n\right)$. Восстановление осуществляется за счет перераспределения функций между ее

модулями. Задачу нахождения новой конфигурации системы можно описать в следующей постановке:

$$\Phi = \sum_{i=1}^{n}\sum_{j=1}^{p} b_{ij}\varphi_{эф}\left(i, j, B\right) = \max\left(\min\right);$$

$$\sum_{i=1}^{n} b_{ij} = 1, \forall j = \overline{1, p};$$

$$\sum_{j=1}^{p} b_{ij} = 1, \forall i = \overline{1, n};$$

$$\Phi \geq \Phi^{*}\left(\Phi \leq \Phi^{*}\right).$$

где Φ^{*} - величина, определяющая минимально (максимально) допустимую эффективность.

Искомое множество S_f включает лишь характеристические векторы, для которых разрешима приведенная задача. На этом этапе решения из множества характеристических векторов состояния инфосюжета S_f выделим подмножество S_q, определяющее состояния системы, в которых возможен выбор конфигураций, обеспечивающих выполнение цели функционирования.

В качестве оценки живучести инфосюжета можно *взять мощность множества* S_q. В случае инфосюжетов на первое место выходит проблема информирования относительно их различных аспектов независимо от наличия или отсутствия неблагоприятных факторов. В связи с этим, в качестве количественного критерия оценки живучести целесообразно использовать отношение количества функций, выполняемых системой при наличии определенных неблагоприятных воздействий либо множества таких воздействий, к общему количеству функций системы, с учетом критичности выполняемых и не выполняемых функций. Критичность каждой конкретной функции определяется индивидуально для каждого конкретного

инфосюжета исходя из его специфики.

Количественный показатель живучести конкретного инфосюжета в заданных условиях можно вычислять по формуле: $S = \sum_{i \in \Delta} \alpha_i / \sum_{j \in \Theta} \alpha_j$, где Θ - множество всех функций информирования, Δ – множество функций инфосюжета, выполняемых в заданных условиях ($\Delta \subseteq \Theta$), α_n – критичность *n*-й функции. Таким образом, количественная оценка живучести инфосюжета будет изменяется в интервале [0, 1], живучесть тем выше, чем больше ее количественная оценка.

Понятие *живучести* системы (инфосюжета) подразумевает ее способность своевременно выполнять свои функции (в рассматриваемом случае информирования) в условиях действия дестабилизирующих факторов. В случае информационных сюжетов такими факторами могут выступать как удаление отдельных документов из информационного пространства, потеря их актуальности, доступности. Привлечение внимания аудитории к другой теме, порождение другого информационного сюжета также может снизить актуальность текущего инфосюжета. Вместе с тем с точки зрения теории живучести происхождение неблагоприятного воздействия играет значительно меньшую роль, чем его последствия.

Живучесть, проявляющаяся как способность целенаправленных инфосюжетов выполнять свои функции информирования на заданном отрезке времени без отказов, определяет минимальный порог устойчивости, за которым без восстановления компонент и функций инфосюжет может потерять свою актуальность и возможность влияния. Вследствие этого и многих других факторов живучесть информационных сюжетов имеет важнейшее значение для информационной безопасности.

2. ИНФОРМАЦИОННЫЙ ПОИСК

Традиционно используемый математический аппарат и инструментальные средства информационного поиска сегодня уже не способны в полной мере удовлетворять потребности пользователей. Изначальная парадигма поисковых систем, сформированная несколько десятилетий тому назад, уже не отвечает реальной ситуации – объемам и динамике информационных потоков, сетевой топологии. Необходим поиск новых принципов, в рамках которых оказалось бы возможным проектирование качественно новых систем обработки больших и динамичных массивов данных.

Перспективы эффективного охвата информационного пространства будут зависеть как от создания и развития эффективной сетевой инфраструктуры, так и развития теоретических основ информатики. В этой связи одной из актуальнейших задач, стоящих перед исследователями различных специальностей, является построение адекватных моделей сетевого информационного пространства и информационного поиска, которые базируются на достижениях в областях лингвистики и информатики, строгом математическом инструментарии.

Предполагается, что должна быть создана теоретическая база для разработки автоматизированных систем мониторинга, адаптивного агрегирования и обобщения информационных потоков, построения и ведения информационных ресурсов сверхбольших объемов и разнообразной тематической направленности. Ожидаемые результаты позволят совместить в единой технологической цепочке мониторинг, информационный поиск, агрегирование информации с содержательным анализом данных, их обобщением, что повысит качество обработки сетевой информации, соответственно, эффективность информационно-аналитической поддержки научно-аналитической деятельности отечественных ученых и специалистов.

В настоящее время структура, объемы и динамика информационного пространства (прежде всего, Интернет-пространства) обусловливают актуальность поисковых технологий. Большинство пользователей Интернет осуществляет поиск информации с помощью сетевых информационно-поисковых систем. Доступ пользователей к современным информационным сетям, эффективное удовлетворение их информационных потребностей возможно только с помощью развитых средств навигации в этих сетях.

Основополагающими характеристиками ИПС являются полнота и точность результатов поиска. Полнота поиска тесно связана с оперативностью охвата информации системой. Созданная однажды база данных Интернет-ресурсов является «слепком» состояния Сети в конкретный момент. Если эта база не будет обновляться постоянно и оперативно, то многие из присутствующих в ней ссылок окажутся «мертвыми». Кроме того, отсутствие оперативности обновления баз данных не позволит пользователю отслеживать последние изменения в его предметной области.

Для пользователей ИПС большое значение имеют также такие характеристики, как скорость обработки запросов, достоверность отклика (например, оцениваемая по источникам), а также дополнительные сервисы – возможность нахождения документов, подобных уже имеющимся, возможность подключения средств автоматического реферирования и перевода и, конечно же, возможность уточнения запроса.

Поисковые машины следующих поколений должны будут лучше классифицировать информацию и нагляднее представлять ее. Поиск не должен ограничиваться лишь обработкой введенных ключевых слов. Кроме того, имеет смысл перехода к концепции смысловой навигации в информационных потоках, как к распределенному во времени интерактивному процессу локализации отдельных семантических секторов в общем информационном потоке. Системы должны будут отслеживать интересы пользователей,

делая поиск более целенаправленным. Новые поисковые машины будут находить опубликованные в сети текстовые, аудио- и видеоматериалы, которые в настоящее время недоступны.

2.1. Проблемы навигации в информационном пространстве

В настоящее время основными проблемами в области информационного поиска являются: небходимость охвата больших объемов информации, большая и сложная динамика информационных потоков, многократное дублирование информации, избыток шумовой информации, спама, наличие скрытого веб-пространства, недоступного современным ИПС, отсутствие реальной модели веб-пространства, эффективных алгоритмов поиска в распределенных (например, пиринговых, социальных) сетях, средств смыслового поиска, поиска мультимедийной информации, мультиязычных средств поиска, отсутствие в свободном доступе универсальных поисковых служб, обеспечивающих поиск фактографической информации, текстовых документов и связей объектов поиска, слабый учет персональных информационных потребностей пользователей, слабая адаптация под эти потребности, явный конфликт при доступе к свободно доступной и/или коммерческой информации. Раскроем некоторые из названных пунктов более подробно.

Необходимость охвата больших объемов информации

В начале существования World-Wide Web (WWW) небольшое количество веб-сайтов публиковало информацию отдельных авторов для относительно большого количества посетителей. Сегодня с появлением и развитием идеологии Web 2.0 ситуация изменилась. Сами посетители веб-сайтов активно участвуют в создании контента, что привело к резкому росту объема и динамики информационного пространства.

Информации в Сети появляется больше, чем ее

успевают охватить поисковые системы. Естественно, это влияет на полноту поиска, что объясняет жесткую конкурентную борьбу за объемы проиндексированных веб-документов, ведущуюся поисковыми службами. С самого начала поисковые системы вели ожесточенную борьбу именно за этот показатель. На первых страницах таких поисковых сайтов, как Altavista, Google, Alltheweb, Yahoo! публиковались соответствующие цифры – количество проиндексированных документов (объем индекса). В начале XXI-го века лидером по охвату ресурсов оказалась служба Google. Однако в 2002 году система Alltheweb неожиданно вышла на первую позицию и была признана лучшей сетевой ИПС в мире по охвату ресурсов, проиндексировав 2,1 млрд. веб-страниц. Затем лидерство вновь вернулось Google – свыше 3,3 млрд. веб-страниц в 2003 году. Последняя цифра, размещенная на титульной странице Google в 2005 г., составляла чуть более 8 млрд. страниц. После этого цифры перестали публиковаться. Из официальных пресс-релизов 2005 г. известно, что объем индекса Google составлял 13 млрд. документов, объем индекса Yahoo! превысил это значение и достиг на то время 20 млрд. документов. Администрация Google была не согласна с этой цифрой, выступая с опровержением. Вместе с тем в заявлении Yahoo! было сказано: "Мы поздравляем Google с изъятием с их главной страницы числа, показывающего размер индекса, и с признанием того, что оно ничего не значит. Как мы уже говорили, важно лишь, чтобы потребители находили то, что они ищут, и мы предлагаем пользователям сравнить результаты поиска наших систем".

Казалось бы, возвращаться к оценке объема индекса никто не будет. Однако в июле 2008 года появилась новая глобальная поисковая система Cuil с относительно небольшим бюджетом (33 млн. долларов), содержащая в индексе 121 млрд. веб-страниц, что, по мнению экспертов, в несколько раз превышало индекс Google, который официально не обнародовался. Можно лишь косвенно сравнивать показатели Google и Cuil, задавая им простейшие запросы (информации Cuil

можно доверять – ее создатели предъявили поисковый индекс внешним экспертам). Как явствует из материалов компаний, обе поисковые системы не используют так называемый стоп-словарь, т.е. запросы по простым, часто употребляемым словам позволят оценить соотношение объемов индексов. И такую оценку с определенным уровнем достоверности может сделать каждый. Например, введя поисковое слово "the" одновременно двум системам, можно получить:

Google: about 22,550,000,000 for the;

Cuil: 22,883,636,124 results for the.

Результаты вполне сопоставимы – можно сделать вывод о примерно одинаковом объеме поисковых индексов. Введем слово "для" (для проверки русскоязычной части), получено:

Google: about 546,000,000 for для;

Cuil: 368,508,113 results for для.

Русскоязычная часть индекса Google оказалась несколько большей. О низком качестве (объеме) русскоязычного индекса Cuil свидетельствуют и запросы по другим словам.

Неожиданный результат получается для еще одного слова – "of":

Google: about 22,760,000,000 for of;

Cuil: 121,000,000,000 results for of.

В этом случае у Cuil результат более чем в 5 раз весомей. Но, учитывая итоги поиска по слову "the" (и по другим словам, в частности, не только на английском языке), можно сделать иной вывод. Каковы бы ни были результаты подобных сравнений, факт остается фактом: Google – самая популярная поисковая система, самый дорогой бренд в мире, а Cuil – мало кому известный проект с бюджетом региональной поисковой системы. Это подтверждает тот факт, что ситуация на рынке поисковых систем не простая – она отражает принцип

новой экономики: здесь не может быть вторых ролей. Или система лучшая в мире, или ею никто не будет пользоваться. Система должна найти свою нишу в задаче максимального удовлетворения запросов пользователей – быть самой полной, самой демократичной, самой интеллектуальной или самой локализованной.

Дублирование информации

Документы, публикуемые на веб-сайтах, зачастую многократно дублируются в виде перепечаток или пересказов. Практически все сетевые ИПС содержат компоненты определения содержательного дублирования. Однако достижение приемлемого качества выявления подобных документов (дубликатов) при использовании различных критериев является актуальной научно-прикладной проблемой. Задача выявления дубликатов, а также перепечаток документов с небольшими изменениями («почти дублей») является одной из актуальнейших и сложнейших при интеграции информационных ресурсов. Существующие в настоящее время алгоритмы выявления дублей в современных информационных потоках требуют применения самых современных компьютерных комплексов, содержащих тысячи серверов (что можно видеть на площадках современных сетевых поисковых служб) и суперкомпьютеров.

Если нахождение явно дублирующейся информации не представляет проблем, то смысловые дубликаты выявляются не так легко, здесь на помощь приходят алгоритмы сопоставления и вероятностной оценки содержимого документов. Кроме того, Интернет является «агрегатором» информации, не находящейся в открытом доступе.

Избыток шумовой информации, спама

Шумовая информация, зачастую несанкционированно навязываемая пользователям (не только электронной почты, но и других сетевых сервисов), в последнее время получила название «спам».

Проблема спама породила две задачи – задачу его выявления и задачу извлечения небольшого количества информации, действительно необходимой пользователю.

Сегодня спам выявляется с помощью сложных комбинированных алгоритмов, требующих как правило мощных вычислительных ресурсов. Вместе с тем, смысловой основой этих алгоритмов является так называемый «наивный» метод Байеса, в рамках которого подразумевается использование оценочной базы – двух текстовых корпусов (например, электронных писем), один из которых составлен из спама, а другой – из полезных документов. Для каждого из текстовых корпусов подсчитывается частота использования каждого слова, после чего вычисляется весовая оценка (от 0 до 1), характеризующая условную вероятность того, что сообщение с этим словом является спамом. Значения весов, близкие к ½, не учитываются при интегрированном расчете, поэтому слова с такими весами игнорируются и удаляются из словарей.

В соответствии с методом, предложенным Полом Грэмом, если сообщение содержит n слов с весовыми оценками $w_1, \ldots, w_n$, то условная вероятность того, что письмо окажется спамом, основанная на данных из оценочных корпусов, вычисляется по формуле:

$$Spm = \Pi w_i / (\Pi w_i + \Pi(1-w_i)).$$

Эта формула обосновывается следующим соображением. Предполагается, что S – событие, заключающееся в том, что письмо – спам, A – событие, заключающееся в том, что письмо содержит слово t. Тогда, в соответствии с формулой Байеса, справедливо:

$$P(S \mid A) = \frac{P(A \mid S)P(S)}{P(A \mid S)P(S) + P(A \mid \overline{S})P(\overline{S})}.$$

Если изначально не известно, является письмо спамом или нет, то на основании опыта предполагается, что $P(\overline{S}) = \lambda P(S)$, на основании чего следует:

$$P(S \mid A) = \frac{P(A \mid S)}{P(A \mid S) + \lambda P(A \mid \overline{S})}.$$

Далее полученная формула обобщается следующим образом. Предполагается, что A_1 и A_2 – это события, заключающиеся в том, что письмо содержит слова t_1 и t_2. При этом вводится допущение, что эти события независимы (именно поэтому метод называется «наивным» байесовским). Условная вероятность того, что письмо, содержащее оба слова (t_1 и t_2) является спамом, равна:

$$P(S \mid A_1 \,\&\, A_2) = \frac{P(A_1 \mid S)P(A_2 \mid S)}{P(A_1 \mid S)P(A_2 \mid S) + \lambda P(A_1 \mid \overline{S})P(A_2 \mid \overline{S})} =$$

$$= \frac{p(t_1)\,p(t_2)}{p(t_1)\,p(t_2) + \lambda(1 - p(t_1))(1 - p(t_2))}.$$

Обобщением этой формулы на случай произвольного количества слов и $\lambda = 1$ является формула П. Грэма. Широкое применяемое в антиспамовских фильтрах находит именно значение $\lambda = 1$. Это допущение упрощает вычисления, но искажает действительность и существенно снижает качество работы соответствующих программ.

Проблемы смыслового поиска

Для пользователя пертинентность, соотношение объема полезной для него информации к общему объему полученной информации, имеет решающее значение. При этом следует учитывать, что формальный запрос к системе является предметом творческого осмысления информационной потребности и ее не всегда точно отражает. Достижение высокой пертинентности – основное поле конкурентной борьбы современных поисковых систем. Именно для удовлетворения информационных потребностей пользователей ИПС сегодня интеллектуализируются – получили широкое практическое применение теории и методы семантических сетей, контент-анализа и глубинного анализа текстов (Text Mining).

Над решением проблемы смыслового, содержательного поиска работают многочисленные коллективы ученых и специалистов во всем мире, в частности, консорциум W3C (The World Wide Web Consortium), где реализуется концепция Семантического Web. Наряду с этой концепцией, революционный прорыв обещает дать более общий подход, а именно Web-2 (www.web2con.com/), который предполагает реализацию концепции семантического Web, включая многоуровневую поддержку метаданных, новые подходы к дизайну и соответствующему инструментарию, технологию глубинного анализа текстов, а также идеологию веб-сервисов, базируясь при этом на информационных ресурсах, накопленных в WWW первого поколения.

Следует признать, что многие основные задачи Семантического Web в настоящее время выглядят достаточно химерными. Вместе с тем частные решения, полученные при попытках реализации Семантического Web, сегодня широко применяются в информационных технологиях. К таким решениям относятся, например, агрегация новостей или ведение блогов (интерактивных сетевых журналов) на основе XML (eXtensible Markup Language — расширяемый язык разметки).

Универсальные поисковые службы

Существующие доступные фактографические базы данных структурированной информации не всегда могут прийти на помощь исследователю-аналитику. Для оперативного определения фактов и сущностей, моделирования информационных связей между ними наиболее перспективным подходом оказывается учет информации, знаний, которые содержатся в неструктурированных текстовых документах, в частности, в Интернет. Поиск в массивах неструктурированной текстовой информации может применяться для задач наведения исследователей-аналитиков «на цель» в условиях, когда фактографические базы данных структурированной информации труднодоступны, неполны, неоперативны.

Неструктурированные тексты содержат в себе несравненно больше важной информации, чем структурированные записи баз данных, именно в силу того, что формализации подлежит сравнительно небольшой сегмент информации. В настоящее время появляется все больше качественных инструментальных средств извлечения понятий из неструктурированных текстов.

Сегодня, когда у пользователей уже накоплен большой опыт работы с традиционными ИПС, оказалось очевидным, что факты или понятия, которые ищутся с помощью таких систем, сами по себе зачастую бессмысленны. Например, если пользователя интересуют информационные связи Сбербанка России с другими банками или частными лицами, то он не знает, какие банки или фамилии ему указать в запросе, а охватить все документы, содержащие словосочетание «Сбербанк России» физически невозможно. В таких случаях информационные связи, интенсивность которых выходит за рамки статистического фона, как правило, отражают реальность.

Интерпретируют обычно не сами понятия или факты, а взаимосвязи между ними. "…важным оказывается не столько исследование самих понятий, сколько исследование их взаимосвязи. Именно взаимосвязь способствует пониманию мотивационно-целевых особенностей отношений человека…" [Массон, 2004] . Т.е. пользователя интересует не понятие само по себе, а понятие в окружении, что позволяет ему сразу иметь представление о предметной области, при необходимости направить уточняющий поиск в нужном направлении. Элементы такого подхода можно видеть, например, в «облаках» системы Quintura (http://quintura.ru), но там отображаются не понятия/сущности, а наиболее часто используемые термы.

Таким образом объективно существует необходимость построения эффективной полнотекстовой ИПС, обеспечивающей поиск не по

отдельным термам или понятиям, а по взаимосвязям между сущностями, присутствующими в документах, то есть создания систем, которые будем условно называть «базами данных связей» (БДС).

В корпоративной информационной инфраструктуре база данных связей может использоваться различным образом, например, отдельно, либо возможности БДС могут быть дополнены возможностями существующих полнотекстовых и/или фактографических баз данных. При этом основным результатом работы БДС является построение карт связей, а в качестве побочного эффекта, реализующего «режим доказательства», может рассматриваться извлечение самих документов как источников связей.

При проектировании БДС должны использоваться решения, которые можно отнести к самым перспективным в области создания информационно-аналитических систем, в частности, теория и технологии глубинного анализа тестов – Text Mining, в том числе развитая методология экстрагирования понятий, теория и технологии баз данных сверхбольших объемов, концепция «сложных сетей». Теория сложных сетей изучает характеристики, учитывая не только топологию сетей, но и статистические феномены, распределение весов отдельных вершин (в качестве которых можно рассматривать сущности, понятия, факты) и ребер, эффекты протекания и проводимости в сетях и т.п.

Анализируя связи в сети, можно определить многие неочевидные свойства, например, выявить наличие кластеров, определить их состав, различия в связности внутри и между кластерами, идентифицировать ключевые элементы, которые связывают кластеры между собой и т.п. Серьезным препятствием при анализе является неполнота информации о связях между отдельными узлами сети. Вместе с тем сегодня уже существуют алгоритмы, с помощью которых становится возможным с высокой вероятностью восстановить отсутствующие фрагменты связей. Даже не имея полного описания

информационной сети, можно получать репрезентативную выборку «реальных» связей и по ней достроить всю сеть.

Учет персональных потребностей пользователей

Представляется очень важным, чтобы агрегирование информации, обеспечение доступа пользователей к этой информации было адаптивным, т.е. ориентированным на информационные потребности конкретных пользователей. Если учитывать динамику и объемы доступной информации в Интернет, то становится очевидным, что обеспечение эффективного доступа в режиме поиска к информации в отрыве от информационных потребностей является практически неразрешимой задачей.

Основная идея адаптивного агрегирования информации заключается в сборе и хранении в информационном хранилище только той информации, которая соответствует информационным потребностям пользователей (существующих или потенциальных). Для этого предполагается, что по мере развития системы в ее информационное хранилище будут попадать актуальные документы из Интернет, соответствующие текущим запросам пользователей. Естественно, с ростом количества пользователей объемы информационного хранилища (репозитория) будут также расти, что в некоторый момент потребует пересмотра его содержания по некоторым критериям, например, по времени в соответствии с формулой Бартона-Кеблера, или по содержанию, используя методы Text Mining.

2.2. Модели информационного поиска

Доступ к современным информационным сетям, эффективное удовлетворение информационных потребностей возможно только с помощью развитых средств содержательного поиска в этих сетях, основным инструментом которого являются информационно-поисковые системы (Retrieval Systems, ИПС), обеспечивающие поиск в гигантских объемах текстовой

информации.

Сегодня миллионам пользователей Интернет известны такие информационно-поисковые системы, как Google, Bing, Yahoo, AltaVista, Яndex, Rambler, которые охватывают миллиарды веб-документов. В основу работы всех подобных систем положены специальные алгоритмы, являющиеся модификациями основных подходов – моделей поиска.

В основу традиционных методов положены три главных подхода, первый из которых базируется на теории множеств (булева модель), второй – на векторной алгебре (векторно-пространственная модель), а третий – на теории вероятностей (вероятностная модель). Эти подходы могут применяться на практике и в каноническом виде, однако у них есть общий недостаток, обусловленный предположением, что содержание документа определяется множеством слов и устойчивых словосочетаний – термов (англ. – Terms), входящих в него без учета взаимосвязей, как «мешок со словами» (англ. – Bag of Words), и, более того, считаются независимыми. Конечно же, такое предположение ведет к потере содержательных оттенков, тем не менее оно позволяет реализовать поиск и группирование документов по формальным признакам. Известны следующие основные недостатки названных моделей:

- Булева модель – невысокая эффективность поиска, отсутствие контекстных операторов, невозможность ранжирования результатов поиска.
- Векторно-пространственная модель связана с расчетом массивов высокой размерности и в каноническом виде малопригодна для обработки больших массивов данных.
- Вероятностная модель характеризуется низкой вычислительной масштабируемостью (т.е. резким снижением эффективности при росте объемов данных), необходимостью постоянного обучения системы.

Системы, построеные на “рафинированных” поисковых моделях, недостаточно оперативны и обладают слабо развитыми поисковыми возможностями и средствами обобщения данных.

Кроме названных, существуют и другие модели поиска, например, семантические, в рамках которых делаются попытки организации смыслового поиска путем анализа грамматики текста, использования баз знаний, тезаурусов, онтологий, которые реализуют семантические связи между отдельными словами и их группами. Вместе с тем, эффективность систем, базирующихся на таких подходах, пока остается невысокой.

Перед рассмотрением отдельных моделей сформулируем некоторые допущения и понятия.

Пусть i – индекс терма t_i из словаря T $(i=1,...,M)$, $d^{(j)}$ – документ, принадлежащий множеству документов D, а $w_i^{(j)} \geq 0$ – вес, ассоциированный с парой $\left(t_i, d^{(j)}\right)$.

Для каждого терма t_i, который не входит в документ $d^{(j)}$, его вес равен нулю: $w_i^{(j)} = 0$. Документ $d^{(j)}$ будем рассматривать как вектор $d^{(j)} = \left(w_1^{(j)}, w_2^{(j)}, ..., w_M^{(j)}\right)$.

Введем также в рассмотрение инверсную функцию g_i, соответствующую индексу терма t_i, которая определяется следующим образом: $g_i\left(d^{(j)}\right) = w_i^{(j)}$.

Классическая булева модель

Булева модель базируется на теории множеств и математической логике. Популярность этой модели связана прежде всего с простотой ее реализации,

которая позволяет индексировать и выполнять поиск в больших документальных массивах.

В рамках булевой модели документы и запросы представляются в виде множества термов – ключевых слов и устойчивых словосочетаний. Каждый терм представлен как булева переменная: 0 (терм из запроса не присутствует в документе) или 1 (терм из запроса присутствует в документе). При этом весовые значения терма в документе принимает лишь два значения: $w_i^{(j)} \in \{0,1\}$.

В булевой модели запрос пользователя представляет собой логическое выражение, в котором термы связываются логическими операторами конъюнкции (AND, $\wedge$), дизъюнкции (OR, $\vee$) и отрицания (NOT, $\neg$). Известно, что любое логическое выражение можно представить дизъюнкцией некоторых выражений, соединенных между собой операцией конъюнкции (дизъюнктивной нормальной формой, ДНФ – DNF).

В булевой модели запрос – это булево выражение, которое приводится к дизъюнктивной нормальной форме, так что его можно записать в виде:

$$q \equiv q_{dnf} = \underset{i=1,..,N}{\vee} q_{cc}^{(i)} ,$$

где $q_{cc}^{(i)}$ – i-я конъюнктивная компонента формы запроса q_{dnf}.

Тогда мера близости документа $d^{(j)}$ и запроса q – $sim\left(d^{(j)}, q\right)$ (от англ. – similarity, близость) в булевой модели определяется выражением:

$$sim\left(d^{(j)}, q\right) = \begin{cases} 1, \text{ если } \exists\ q_{cc}^{(i)}: \quad \left(q_{cc}^{(i)} \in q_{def}\right) \wedge \left(\forall k,\ g_k\left(q_{cc}^{(i)}\right) = g_k\left(d^{(j)}\right)\right), \\ 0, \text{ иначе.} \end{cases}$$

Таким образом $sim\left(d^{(j)},q\right)$ принимает значение 1, если существует такая конъюнктивная компонента $q_{cc}^{(i)}$, входящая в дизъюнктивную нормальную форму q_{dnf}, что инверсная функция каждого терма k данной конъюнктивной компоненты совпадает с этой же инверсной функцией для документа $d^{(j)}$. В противном случае $sim\left(d^{(j)},q\right)$ оказывается равной 0.

Таким образом, если $sim\left(d^{(j)},q\right)=1$, то в соответствии с булевой моделью документ $d^{(j)}$ считается релевантным (соответствующим) запросу q. В противном случае документ не является релевантным. Весовых различий, необходимых для ранжирования документов по уровню соответствия запросу в булевой модели не предусмотрено, что является ее существенным недостатком.

Векторно-пространственная модель поиска

Многие из известных информационно-поисковых систем базируются на векторно-пространственной модели описания данных (Vector Space Model), предложенной Г. Солтоном в 1975 г. и впервые примененной в системе SMART. Данная модель является классической алгебраической. В рамках этой модели документ описывается вектором в евклидовом пространстве, в котором каждому терму, использующемуся в документе, ставится в соответствие его весовое значение, определяемое на основе статистической информации о его появлении как в отдельном документе, так и во всем документальном массиве. Описание запроса, соответствующего необходимой пользователю тематике, также представляет собой вектор в том же евклидовом пространстве термов. Для оценки близости запроса и документа используется скалярное произведение соответствующих векторов запроса и документа.

В рамках этой модели каждому терму t_i в документе $d^{(j)}$ соответствует некоторый неотрицательный вес $w_i^{(j)}$.

Каждому запросу q, который представляет собой также множество термов, не соединенных между собой никакими логическими операторами, также соответствует вектор весовых значений w_i^q.

Таким образом, каждый документ и запрос могут быть представлены в виде n-мерного вектора, где n – общее количество термов в словаре модели. В соответствии с рассматриваемой моделью, близость документа $d^{(j)}$ к запросу q, которые, как и в предыдущих моделях, рассматриваются как информационные векторы $\boldsymbol{d}_j = (w_1^{(j)}, w_2^{(j)}, ..., w_n^{(j)})$ и $\boldsymbol{q} = (w_1^q, w_2^q, ..., w_n^q)$, оценивается как их скалярное произведение. При этом вес отдельных термов можно вычислять разными способами. Один из возможных простейших подходов – использовать как вес терма $w_i^{(j)}$ в документе нормализованную частоту $freq_i^{(j)}$ его встречаемости в данном документе, то есть:

$$w_i^{(j)} = freq_i^{(j)} / \max_{1 \le k \le n} freq_k^{(j)}.$$

Вычисленный таким образом вес терма в документе принято обозначать аббревиатурой $tf_i^{(j)}$ или TF (англ. – Term Frequency – частота термина).

Однако этот подход не учитывает, насколько часто рассматриваемый терм используется во всем массиве документов, так называемую, дискриминационную силу терма. Поэтому в случае, когда доступна статистика использования термов во всем документальном массиве, более эффективно следующее правило вычисления веса:

$$w_i^{(j)} = tf_i^{(j)} \cdot \log \frac{N}{n_i},$$

где n_i – количество документов, в которых используется терм t_i, а N – общее количество документов в массиве. Например, если некоторое слово встречается в каждом документе массива, то его использование в запросе, очевидно, бесполезно. Соответственно, в этом случае $n_i = N$, и, следовательно, $w_i^{(j)} = tf_i^{(j)} \cdot \log \frac{N}{N} = 0.$

Следует отметить, что приведенная выше формула многократно уточнялась с целью наиболее точного соответствия выдаваемых документов запросам пользователей. В 1988 году Солтоном был предложен следующий вариант для вычисления веса терма t_i из запроса в документе:

$$w_i^q = \left(0.5 + \frac{freq_i^q}{\max\limits_{1 \le l \le n} freq_l^q} \right) \cdot \log \frac{N}{n_i},$$

где $freq_i^q$ – частота терма t_i из запроса в тексте этого документа.

Обычно весовые значения $w_i^{(j)}$ нормируются путем деления на их общую сумму. Такой метод взвешивания термов имеет стандартное обозначение $TF \cdot IDF$, где TF указывает на частоту появления терма в документе, а IDF – на величину, обратную количеству документов в массиве, содержащих данный терм (от англ. – Inverse Document Frequency).

Когда возникает задача определения тематической близости двух документов или документа и запроса, в этой модели используется простое скалярное произведение $sim(d^{(1)}, d^{(2)})$, двух соответствующих векторов весовых значений $(w_1^{(1)}, w_2^{(1)}, ..., w_n^{(1)})$ и

$(w_1^{(2)}, w_2^{(2)},..., w_n^{(2)})$, которое соответствует косинусу угла между векторами – образами документов $\boldsymbol{d}^{(1)}$ и $\boldsymbol{d}^{(2)}$. Очевидно, $sim(\boldsymbol{d}^{(1)}, \boldsymbol{d}^{(2)})$ принадлежит диапазону [0, 1]. Чем больше величина $sim(\boldsymbol{d}^{(1)}, \boldsymbol{d}^{(2)})$, тем более близки документы $\boldsymbol{d}^{(1)}$ и $\boldsymbol{d}^{(2)}$. Для любого документа $\boldsymbol{d}$ имеем $sim(\boldsymbol{d}, \boldsymbol{d}) = 1$. Аналогично мерой близости документа $\boldsymbol{d}^{(j)}$ и запроса $\boldsymbol{q}$ является величина:

$$sim(d_j, q) = \frac{\boldsymbol{d}^{(j)} \cdot \boldsymbol{q}}{|\boldsymbol{d}^{(j)}| \cdot |\boldsymbol{q}|} = \frac{\sum_{i=1}^{n} w_i^{(j)} w_i^q}{\sqrt{\sum_{i=1}^{n} (w_i^{(j)})^2} \sqrt{\sum_{i=1}^{n} (w_i^q)^2}}.$$

Векторно-пространственная модель представления данных обеспечивает системам, построенным на ее основе, такие возможности, как:

- обработку запросов без ограничений их длины;
- простоту реализации режима поиска подобных документов (каждый документ может расматриваться как запрос);
- сохранение результатов поиска с возможностью выполнения уточняющего поиска.

Вместе с тем в векторно-пространственной модели не предусмотрено использование логических операций в запросах, что существенно ограничивает ее применимость.

Вероятностная модель поиска

В 1977 году С. Э. Робертсон и К. Спарк-Джонс обосновали и реализовали вероятностную модель поиска. В данной модели поиска вероятность того, что документ релевантен запросу, основывается на предположении, что термы запроса по-разному

распределены среди релевантных и нерелевантных документов. При этом используются формулы расчета вероятности, базирующиеся на теореме Байеса.

Основной вопрос, который решается с помощью модели: как велика вероятность того, что документ d релевантен запросу q? Релевантность при этом рассматривается как вероятность того, что данный документ может оказаться интересным пользователю. Функционирование модели базируется как на экспертных оценках, получаемых в результате обучения модели, которые признают документы из учебной коллекции релевантными/нерелевантными, так и на последующих оценках вероятности того, что документ является релевантным запросу исходя из состава его термов.

Если для запроса известны эти оценки вероятностей для всех документов, то документы можно сортировать по ним и выводить пользователям в нисходящем порядке. То есть вероятностная модель поиска предусматривает определение вероятностей соответствия запросу для документов, сортировку и предоставление документов с ненулевой вероятностью пользователю.

С самого начала в вероятностной модели использовалось упрощение, которое допускает независимость вхождения в документ любой пары термов (поэтому такой подход называется «наивным» байесовским).

При этом в вероятностной модели поиска предполагается наличие учебных наборов релевантных и нерелевантных документов, выбранных пользователем или полученных автоматически при каком-то начальном предположении. Вероятность того, что поступивший документ является релевантным, рассчитывается на основании соотношения появления термов в релевантном и нерелевантном массиве документов.

Рассмотрим основу модели, а именно байесовский

подход, более детально. Пусть X, Y – два независимых события, X, $Y \subset G$, G – базовое вероятностное пространство.

Вероятность X при условии Y определяется таким образом:

$$P(X \mid Y) = \frac{P(X \bigcap Y)}{P(Y)}.$$

Известно, что из этого соотношения следует формула Байеса:

$$P(X \mid Y) = \frac{P(X \bigcap Y)}{P(Y)}; \;\; P(Y \mid X) = \frac{P(Y \bigcap X)}{P(X)};$$

$$P(Y \bigcap X) = P(X \bigcap Y) \Rightarrow P(X \mid Y) = \frac{P(Y \mid X) \cdot P(X)}{P(Y)}.$$

Рассмотрим условные вероятности двух событий, а именно того, что документ релевантен (R) запросу – $P(R \mid q,d)$, где q – запрос, d – документ, а также того, что документ нерелевантен ($\bar{R}$) запросу – $P(\bar{R} \mid q,d)$.

Введем понятие квоты релевантности как меры близости документа запросу – $O(R)$:

$$O(R) = \frac{P(R)}{P(\bar{R})} = \frac{P(R)}{1 - P(R)}.$$

Очевидно, что квота меньше, чем 1 для вероятности $P(R) < 0.5$ и больше 1 для вероятности $P(R) > 0.5$.

Определим квоту того события, что документ релевантен запросу:

$$O\left(R \mid q,d\right) = \frac{P(R \mid q,d)}{P(\bar{R} \mid q,d)}.$$

Для числителя этой формулы справедливо:

$$P(R \mid d,q) = \frac{P(R \cap q \cap d)}{P(q \cap d)} =$$
$$= \frac{P(d \mid R \cap q) \cdot P(R \cap q)}{P(d \mid q) \cdot P(q)} = \frac{P(d \mid R \cap q) \cdot P(R \mid q)}{P(d \mid q)}.$$

Подставляя соответствующие выражения в числитель и знаменатель, получаем формулы для квоты релевантности:

$$O(R \mid d,q) = \frac{P(R \mid q,d)}{P(\bar{R} \mid q,d)} =$$
$$= \frac{\frac{p(d \mid R \cap q) \cdot p(R \mid q)}{p(d \mid q)}}{\frac{p(d \mid \bar{R} \cap q) \cdot p(\bar{R} \mid q)}{p(d \mid q)}} = \frac{p(R \mid q)}{p(\bar{R} \mid q)} \times \frac{p(d \mid R \cap q)}{p(d \mid \bar{R} \cap q)}.$$

Перейдем к рассмотрению документа как вектора термов. Пусть $T = \{t_1, ..., t_n\}$ – множество термов, которые содержатся в корпусе документов *D*. Документ рассматривается как вектор из бинарных значений весов входящих в него термов $\vec{d} = (w_1, ..., w_n)$, где:

$$w_i = \begin{cases} 1, & t_i \in d; \\ 0, & t_i \notin d. \end{cases}$$

Тогда:

$$p(d \mid R,q) = p(\vec{d} \mid R,q) = \prod_{i=1}^{n} p(t_i \mid R,q).$$

Последнее вытекает из упрощения, связанного с предполагаемой независимостью термов. В результате квота релевантности принимает вид:

$$O(R \mid q,d) = \frac{p(R \mid q)}{p(\bar{R} \mid q)} \cdot \frac{p(d \mid R \cap q)}{p(d \mid \bar{R} \cap q)} = O(R \mid q) \cdot \prod_{i=1}^{n} \frac{p(t_i \mid R \cap q)}{p(t_i \mid \bar{R} \cap q)}.$$

Модель предусматривает еще одно упрощение, а именно то, что при $t_i \in T \setminus q : p(t_i \mid R,q) = p(t_i \mid \bar{R},q)$.

Разложим произведение следующим образом:

$$O(R \mid q,d) = O(R \mid q) \times$$
$$\times \prod_{t_i \in q \cap d} \frac{p(t_i \mid R \cap q)}{p(t_i \mid \bar{R} \cap q)} \cdot \prod_{t_i \in q \setminus d} \frac{p(t_i \mid R \cap q)}{p(t_i \mid \bar{R} \cap q)} \cdot \prod_{t_i \notin q} \frac{p(t_i \mid R \cap q)}{p(t_i \mid \bar{R} \cap q)}.$$

Последний сомножитель равен единице ввиду вышеприведенного предположения. Введем обозначение:

$$r_i = p(w_i = 1 \mid R,q);$$
$$n_i = p(w_i = 1 \mid \bar{R},q).$$

В этих обозначениях выполняется:

$$O(R \mid q,d) = O(R \mid q) \cdot \prod_{t_i \in q \cap d} \frac{r_i}{n_i} \cdot \prod_{t_i \in q \setminus d} \frac{1-r_i}{1-n_i}.$$

Учитывая то, что:

$$\prod_{t_i \in q \cap d} \frac{(1-r_i)(1-n_i)}{(1-n_i)(1-r_i)} = 1,$$

получаем:

$$O(R \mid q,d) = O(R \mid q) \cdot \prod_{t_i \in q \cap d} \frac{r_i(1-n_i)}{n_i(1-r_i)} \cdot \prod_{t_i \in q} \frac{1-r_i}{1-n_i}.$$

Для исследования релевантной последовательности элементов достаточно учитывать только второй сомножитель, при этом его можно прологарифмировать (логарифм – монотонная функция, которая не меняет рангов документов). То есть можно анализировать сумму:

$$\sum_{t_i \in q \cap d} \log \frac{r_i(1-n_i)}{n_i(1-r_i)} = \sum_{t_i \in q \cap d} \left[\log \frac{r_i}{n_i} + \log \frac{1-n_i}{1-r_i} \right].$$

Рассмотрим приближенные значения, полученные на базе анализа некоторой предварительно полученной обучающей выборки:

$$\tilde{r}_i = \frac{rel_i}{rel}; \quad \tilde{n}_i = \frac{nrel_i}{nrel},$$

где rel_i – количество релевантных документов, которое содержит терм с индексом i; $nrel_i$ – количество нерелевантных документов, которые содержат терм с индексом i.

То есть можно анализировать сумму, называемую поисковым статусом:

$$SV_i = \sum_{t_i \in q \cap d} \log \frac{\tilde{r}_i(1-\tilde{n}_i)}{\tilde{n}_i(1-\tilde{r}_i)} = \sum_{t_i \in q \cap d} \log \frac{rel_i(nrel - nrel_i)}{nrel_i(rel - rel_i)}.$$

2.3. Информационно-поисковые системы

Доступ пользователей к современным информационным сетям, эффективное удовлетворение их информационных потребностей возможны только с помощью развитых средств навигации в этих сетях. Основным инструментом при этом выступают информационно-поисковые системы, обеспечивающие поиск в гигантских объемах текстовой информации.

Существующие доступные фактографические базы данных структурированной информации не всегда могут прийти на помощь исследователю-аналитику. Для оперативного определения фактов и сущностей, моделирования информационных связей между ними наиболее перспективным подходом оказывается учет информации, знаний, которые содержатся в неструктурированных текстовых документах, в частности, в Интернет. Поиск в массивах неструктурированной текстовой информации может

применяться для задач наведения исследователей-аналитиков «на цель» в условиях, когда фактографические базы данных структурированной информации труднодоступны, неполны, неоперативны.

Первые реально функционирующие полнотекстовые информационно-поисковые системы появились в начале компьютерной эры. Назначением этих систем был поиск в библиотечных каталогах, архивах, массивах документов, таких как статьи, нормативные акты, рефераты, брошюры, диссертации, монографии.

Основными функциями информационно-поисковых систем изначально были:

- хранение больших объемов информации;
- быстрый поиск необходимой информации;
- добавление, удаление и изменение хранимой информации;
- вывод информации в удобном для пользователя виде.

В 1966 году 16-ю американскими библиотеками для установления стандартного формата для электронных каталогов была начата реализация проекта MARC (http://www.loc.gov/marc/), обеспечившего переход к унифицированному обмену электронными данными, что способствовало эффективной организации электронных каталогов. Внедрение стандартного библиографического формата позволило библиотекам объединить усилия. В 1972 году получил международное признание стандарт MARC-2, на основе которого были созданы многие национальные стандарты.

В начале 1970-х годов коммерческие компьютерные службы уже предоставляли возможность интерактивного поиска в тематических базах данных Национальной медицинской библиотеки и Министерства образования США. При этом некоторые

из этих служб существуют и сегодня: основанная еще в 1965 году система Dialog (http://www.dialog.com/), входящая в настоящее время в корпорацию Thomson, сегодня обеспечивает своим клиентам доступ к сотням базам данных.

В начале 1990-х годов для унификации информационных систем был разработан международный стандарт Z39.50 - информационно-поисковый протокол для библиографических систем. В 1994 г. университет Джорджии запустил пилотный проект "Галилей" (http://www.usg.edu/galileo/) с использованием Site-Search - пакета программ Огайского центра, соответствующий стандарту Z39.50. Стандарт Z39.50 также был положен в основу исторически первой службы поиска распределенной информации в Интернет - WAIS (Wide Area Information Service), в настоящее время уже утратившей свою актуальность.

В настоящее время информационные ресурсы только веб-пространства составляют свыше триллиона документов, к которым возможен свободный доступ любого пользователя. Естественно, для того, чтобы найти необходимую информацию и этой крупнейшей распределенной полнотекстовой базе данных необходимо использовать самые мощные ИПС. Такие системы существуют и конкурируют друг с другом. Сегодня миллионам пользователей Интернет известны такие информационно-поисковые системы, как Google, Yahoo, AltaVista, Bing, Яndex, Rambler, которые охватывают миллиарды веб-документов.

Основополагающими характеристиками ИПС являются полнота и релевантность результатов поиска. Полнота поиска тесно связана с оперативностью охвата информации системой. Созданная однажды база данных Интернет-ресурсов является "слепком" состояния Сети в конкретный момент. Если эта база не будет обновляться постоянно и оперативно, то многие из присутствующих в ней ссылок окажутся «мертвыми». Кроме того, отсутствие оперативности обновления баз данных не позволит пользователю отслеживать последние изменения в его предметной области.

Для пользователей ИПС большое значение имеют также такие характеристики, как скорость обработки запросов, достоверность отклика (например, оцениваемая по источникам), а также дополнительные сервисы - возможность нахождения документов, подобных уже имеющимся, возможность подключения средств автоматического реферирования и перевода и, конечно же, возможность уточнения запроса.

Поисковые машины следующих поколений должны будут лучше классифицировать информацию и нагляднее представлять ее. Поиск не должен ограничиваться лишь обработкой введенных ключевых слов. Кроме того, имеет смысл перехода к концепции смысловой навигации в информационных потоках, как к распределенному во времени интерактивному процессу локализации отдельных семантических секторов в общем информационном потоке. Системы должны будут отслеживать интересы пользователей, делая поиск более целенаправленным. Новые поисковые машины будут находить опубликованные в сети текстовые, аудио- и видеоматериалы, которые в настоящее время недоступны.

2.4. Метапоисковые системы

Несколько больших поисковых систем в Интернете, например Google, Bing, Яндекс не в состоянии проиндексировать все веб-пространство и обеспечить возможность поиска абсолютно всех веб-страниц. Известно, что эти глобальные поисковые системы страдают от ряда ограничений. Например, охват Интернета каждой из них ограничен в силу разных причин, обусловленных топологией сети, ограничениями на глубину обхода отдельных веб-сайтов, временем, необходимым для индексирования той или иной части веб-пространства. Рост поисковых систем приводит к тому, что большая доля информации, охватываемой ими, устаревает. Отрытым остается вопрос о масштабируемости на все веб-пространство современных технологий поиска.

Как дополнения к глобальным поисковым сервисам в настоящее время рассматриваются метапоисковые системы (мультипотоковые поисковые системы) – поисковые инструменты, которые посылают запросы пользователей одновременно на несколько поисковых серверов и, иногда, в так называемый глубинный веб. Собрав результаты, метапоисковая система удаляет дублирующиеся ссылки и, в соответствии со своим алгоритмом, объединяет/ранжует результаты для предоставления их в общем списке. В отличие от других поисковых систем, метапоисковые системы могут не иметь собственных баз данных, содержащих индексированный контент веб-пространства.

Итак, метапоиск – это процесс, который поддерживает унифицированный доступ к нескольким поисковым системам. При этом метапоисковые системы могут не иметь собственного индекса веб-страниц, но в их поисковых механизмах учитываются особенности каждой подключенной поисковой системы. Когда метапоисковая система получает пользовательский запрос, она сначала передает запрос соответствующим поисковым системам, сервис которых она эксплуатирует, а затем собирает (реорганизует, обобщает) результаты используемых поисковых систем. Явным преимуществом метапоисковых систем является возможность поддержки индексных данных в актуальном состоянии, так как каждая локальная поисковая система охватывает собственный фрагмент веб-пространства. Кроме того, для работы метапоисковых систем требуются намного меньшие инвестиции в аппаратные средства по сравнению традиционными поисковыми системами типа Google, в инфраструктуре которых используются сегодня тысячи серверов.

2.4.1. Функционирование метапоисковых систем

Для выбора подключенной поисковой системы в большинстве реализаций метапоисковых систем используется ранжирование баз данных этих систем для конкретного заданного запроса, основанное на

определенных принципах. Например, система gGlOSS использует сумму коэффициентов подобия документов, превышающую определенный порог [Liu, 2000], Система CORI Net использует вероятность того, что база данных системы содержит соответствующие документы, удовлетворяющие условиями данного запроса [Hawking, 1999], Система D-WISE использует сумму взвешенных частот документов, соответствующих условиям запроса. Все эти методы ранжирования базы данных систем не предназначены для получения эффективных результатов поиска на основе некоторых критериев оптимальности. Используемая функция для ранжирования базы данных базируется на сходстве запроса и аналогичного документа в базе данных подключенной поисковой системы.

Для слияния результатов поиска в большинстве существующих подходов используется взвешенное распределение полученных документов путем представления релевантных запросу документов из баз данных поисковых систем с наивысшими рангами (например, CORI Net, D-WISE, ProFusion [Lawrence, 1998]), или откорректированные локальные индексы подобия документов (например, D-WISE, ProFusion). Эвристические подходы не направлены на обеспечение поиска всех потенциально полезных документов для данного запроса [Lui, 2001]. Метапоисковый механизм Inquirus [Lawrence, 1998] использует реальные глобальные индексы подобия документов для их слияния. Чтобы определить, что необходимо вывести документы из локальной базы данных подключенной поисковой системы, решается задача нахождения необходимого локального порога подобия на основе глобального порога подобия. Этот подход направлен на обеспечение поиска всех потенциально полезных документов от каждой подключенной поисковой системы и минимизацию выдачи ненужных документов. Основная проблема таких подходов связана с тем, что локальное сходство – это собственная функция каждой подключенной поисковой системы, а функция подобия, как правило, относится к самой

метапоисковой системе.

2.4.2. Проблема метапоиска документальной информации

Существует несколько серьезных проблем в реализации эффективных метапоисковых систем, среди которых можно назвать, например, проблему выбора поисковых систем, подключаемых по запросу пользователя, т.е. поисковых систем, которые могут содержать релевантные пользовательскому запросу документы. Цель выбора подключаемых поисковых систем – повышение эффективности поиска путем отправления запросов только потенциально полезным поисковым системам. Следующая проблема связана с объединением (слиянием) результатов поиска, ранжированием этих результатов, обеспечением того, чтобы более полезные для пользователя документы шли впереди остальных. Ввиду большой неоднородности поисковых систем, часто трудно добиться хорошего слияния результатов поиска. Эффективный метапоисковый механизм должен обеспечивать иллюзию того, документы находятся в одной базе данных, обеспечивать минимизацию времени поиска.

При проектировании метапоисковых систем решается ряд проблем. Прежде всего, из полученного от подключенных поисковых систем огромного количества документов необходимо выделить наиболее релевантные, т.е. документы, соответствующие запросу пользователя. Как правило, разработчики метапоисковых систем не совсем оправданно надеются, что подключаемые поисковые системы возвращают релевантные результаты поиска, и слишком полагаются на ранг (позицию), который в каждой поисковой системе приписывается документу.

В метапоисковых системах, в которых не проводится анализ полученных описаний найденных документов, могут отображаться нерелевантные документы, которые ошибочно идут первыми в

подключенной поисковой системе, чем существенно снижается качество поиска.

Большинство из существующих метапоисковых систем обеспечивают подключение не более чем к сотням поисковым системам. Зачастую используемые в них подходы не могут масштабироваться до десятков тысяч или больше локальных поисковых систем, обеспечивая высокую эффективность. Проблемы заключаются в первую очередь в выборе наиболее релевантных поисковых систем, подключаемых для обработки конкретного запроса. Это всегда ведет к большим затратам на вычисления и необходимости иметь репрезентативные «семантические шаблоны» для каждой из баз данных.

2.4.3. Типы метапоисковых систем

Существует четыре вида метапоисковых систем [Базак, 2003]:

- «Реальные» метапоисковые системы, которые объединяют/ранжируют результаты на одной странице;
- «Псевдометапоисковые» системы первого типа, которые группируют результаты для каждой поисковой системы и выдают их на одной длинной странице;
- «Псевдометапоисковые» системы второго типа, которые открывают для каждой используемой поисковой системы новое окно;
- Поисковые утилиты – программные поисковые средства.

Рассмотрим подробнее каждый из видов метапоисковых систем и приведем примеры:

1. «Реальные» метапоисковые системы одновременно проводят поиск в основных поисковых системах, обобщают результаты, удаляют дублированные ссылки и предоставляют наиболее подходящие результаты в соответствии с алгоритмом. Примерами таких метапоисковых системам являются:

Nigma (http://www.nigma.ru/), IBoogie (http://iboogie.com);

2. «Псевдометапоисковые» системы первого типа отправляют запрос на подключенные поисковые системы, а потом предоставляют сгруппированнные по поисковым системам результаты в один длинный, но легкий для чтения список с возможностью прокрутки. В зависимости от того, сколько пользователь выбирает поисковых систем, зависит и время ожидания ответа. Такими метапоисковими системами являются, например: Search Wiz (http://www.searchwiz.com/) и Search Fid (http://find.copernic.com/).

3. Существует два типа «псевдометапоисковых» систем второго типа:

- пользователь вводит запрос один раз, а потом выбирает поисковые системы. Для каждой выбранной системы будет открыто новое окно браузера. Такими метапоисковыми системами являются: Multi-Search-Engine (http://www.dogpile.com/), The Info (http://www.theinfo.com/);
- пользователь выбирает поисковую систему, вводит запрос в формуляре поисковой системы, и тогда открывается новое окно. У каждой поисковой системы свой бланк запроса. Среди таких метапоисковых систем можно назвать: Alpha Seek (http://seekingalpha.com) и Freeality (http://www.freeality.com/).

4. Поисковые утилиты (также называемые поисковыми дополнениями рабочего стола) – это загружаемые инструменты метапоиска, которые ищут в многочисленных поисковых системах. Результаты упорядочиваются и ранжируются по релевантности, при этом повторы удаляются. Для большинства из таких систем есть бесплатная пробная версия. Самые популярные из них: Copernic (http://www.copernic.com) и WebFerret (http://www.webferret.com).

2.5. Модели и технологии децентрализированного поиска

2.5.1. Поиск в P2P (пиринговых) сетях

В настоящее время веб-пространство не является крупнейшим информационным ресурсом в Интернете. Основной объем ресурсов сосредоточен в "пиринговых" сетях (P2P – «точка-точка»), многие из которых являются так называемыми «файлообменными». В таких сетях отсутствуют выделенные серверы, а каждый узел является как клиентом, так и сервером. Пиринговые сети состоят из узлов, каждый из которых взаимодействует лишь с некоторым подмножеством других узлов. При освещении этой тематики учитывалось то, что проблемы поиска и уязвимости в таких сетях до сих пор остаются открытыми.

Существует несколько областей применения пиринговых сетей, объясняющих их растущую популярность, назовем некоторые из них:

- Обмен файлами. Сети P2P выступают альтернативой FTP-архивам, которые утрачивают перспективу из-за значительных информационных перегрузок.
- Распределенные вычисления. Например, такой проект с элементами P2P, как SETI@HOME, посвященный распределенному поиску внеземных цивилизаций, продемонстрировал высокий вычислительный потенциал для распараллеливаемых задач. Вместе с тем, этому проекту свойственна централизованная раздача и прием данных.
- Обмен сообщениями. Как известно, ICQ – это P2P-проект. Эта сеть также обладает элементами централизации, в частности, очень зависит от состояния сервера login.icq.com.
- Интернет телефония. Сегодня одной из самых популярных служб Интернет-телефонии

является Skype (www.skype.com), созданная в 2003 г. Н. Зеннстромом и Я. Фриисом, авторами известной пиринговой сети KaZaA. Построенная в архитектуре P2P служба Skype охватывает свыше 10 млн. пользователей.

- Групповая работа. Сегодня реализованы такие сети групповой работы, как Groove Network (защищенное пространство для коммуникаций) и OpenCola (поиск информации и обмен ссылками).

Вопрос эффективного поиска в таких сетях остается открытым, существуют лишь специальные поисковые сайты в веб-пространстве, помогающие решить эту проблему.

На практике пиринговые сети состоят из рабочих станций, каждая из которых взаимодействует лишь с некоторым подмножеством узлов сети (из-за ограниченности ресурсов). Достаточно часто пиринговые сети дополняются выделенными серверами. Такие серверы позволяют решать вопросы поиска по запросам, так как именно эта проблема для пиринговых сетей не может считаться решенной.

Файлообменные P2P-сети уже в начале 2010 г. охватывали более 150 млн. узлов. Сегодня в Интернет более половины всего трафика приходится на файлообменные P2P-сети. Наиболее популярные из них – это Bittorrent, Gnutella2 и eDonkey2000.

При поиске в пиринговых сетях тема полноты поиска отодвигается на второй план, главная же задача – быстрое и эффективное нахождение наиболее релевантных откликов на запрос, передаваемый от рабочей станции всей сети. В частности, актуальная проблема – уменьшение сетевого трафика, порождаемого запросом (например, пересылки запроса по многочисленным рабочим станциям), и в то же время получение наилучших характеристик выдаваемых документов, т.е. получение качественного результата.

Приемлемое качество поиска в пиринговых сетях на сегодняшний день обеспечивают лишь специализированные, централизованно наполняемые, поисковые веб-сайты, работающие по протоколу HTTP. Например, для файлообменной сети eMule таким поисковым сервером является сайт Figator.com, а для сети Bittorrent – сайт isoHunt.com.

Как и для файлообменных сетей, для этих серверов особо актуальными и критичными являются проблемы качества и достоверности предоставляемого контента, фальсификация файлов и распространение фальшивых ресурсов, вирусов, "троянских коней", возможность фальсификации ID рабочих станций.

Существует несколько алгоритмов поиска в таких сетях, ни один из которых не подходит для получения результатов, сравнимых даже с традиционным поиском в веб-пространстве. Наиболее популярные алгоритмы базируются на поиске ресурсов по ключам. В большинстве пиринговых сетей, ориентированных на обмен файлами, используются два вида сущностей, которым приписываются соответствующие идентификаторы (ID): узлы и ресурсы (например, файлы), которые характеризуются ключами (Key), то есть сеть может быть представлена двумерной матрицей размерностью MN, где M – количество узлов, N – количество ресурсов. В этом случае задание поиска сводится к нахождению ID узла, на котором сохраняется ключ ресурса. Одним из наиболее эффективных алгоритмов поиска в сетях P2P является так называемый «Интеллектуальный поисковый механизм» (*Intelligent Search Mechanism, ISM*). Улучшение скорости и эффективности поиска информации с помощью данного метода достигается за счет минимизации расходов на количество сообщений, которые передаются между узлами, а также количества узлов, которые опрашиваются для каждого запроса. То есть оцениваются лишь те узлы, которые больше всего отвечают конкретному запросу.

Поисковый механизм ISM состоит из двух

компонент – профайла и способа его ранжирования, так называемого ранга релевантности. Каждый узел сети строит информационный профайл для каждого из соседних узлов. Профайл содержит последние ответы от каждого из узлов. С помощью ранга релевантности осуществляется ранжирование профайлов узлов для выбора тех соседних, которые будут давать наиболее релевантные документы по запросу.

При реализации модели ISM применяется единый стек запросов, в котором сохраняется по T запросов для N узлов. Как только стек заполняется, происходит замена того запроса, который не использовался дольше (*Least Recently Used*, *LRU*), с целью сохранения последних запросов. Функция «ранг релевантности» (*Relevance Rank, RR*) применяется узлом P_l, чтобы выполнять оперативную классификацию его соседей для определения тех из них, которые стоит опрашивать первыми по запросу q. Для вычисления ранга релевантности каждого узла P_i, P_l сравнивает q со всеми запросами в структуре профайла, для которого известен список ответов на предыдущие запросы, и вычисляет $RR(P_i, q)$:

$$RR(P_i, q) = \sum_{j \in Q} Sim(q_j, q)^{\alpha} \cdot S(P_i, q_j),$$

где α – параметр, который задает вес запросов. В этой формуле Q – множественное число запросов, на которые был ответ от узла P_i; $S\left(P_i, q_j\right)$ – количество результатов, которые возвращались узлом P_i по запросу q_j; метрика Sim рассчитывается по правилу, принятому в векторно-пространственной модели поиска:

$$Sim(q_j, q) = \frac{q_j \cdot q}{\left|q_j\right|\left|q\right|}.$$

Ранг релевантности *RR* обеспечивает более высокий ранг узлу, который возвращает больше результатов.

Метод ISM эффективно работает в сетях, где узлы содержат некоторые специализированные сведения. В частности, исследование сети Gnutella показывает, что качество поиска очень зависит от «окружения» узла, с которого задается запрос. Большая проблема в методе ISM заключается в том, что поисковые сообщения могут циклически проходить те же узлы сети, не достигая некоторых ее частей. Чтобы решить эту проблему для охвата большей части сети предложен подход, при котором для каждого запроса выбиралось небольшое подмножество случайных узлов, которые добавлялись к набору релевантных узлов.

Существуют также другие подходы к решению этой проблемы, например, применяемый в протоколе BGP4 (RFC 1771), где каждый запрос хранит «историю» – список узлов, через которые он уже прошел.

2.5.2. Системы поиска в пиринговых сетях

Сегодня значительный объем информационных ресурсов сети Интернет сосредоточен на называемых пиринговых, файлообменных сетях. На практике пиринговые сети состоят из рабочих станций, каждая из которых взаимодействует лишь с некоторым подмножеством узлов сети (из-за ограниченности ресурсов). Для реализации протокола P2P используются клиентские программы, обеспечивающие функциональность как отдельных рабочих станций, так и всей пиринговой сети в целом.

Достаточно часто пиринговые сети дополняются выделенными серверами. Чаще всего именно такие серверы позволяют решать вопросы поиска по запросам, так как именно эта проблема для пиринговых сетей не может считаться решенной. Вопрос эффективного поиска в таких сетях средствами самих сетей остается пока открытым (поскольку большинство из них не предполагает жесткой централизации, а напротив, они

по определению являются децентрализованными), однако существуют специальные поисковые сайты в веб-пространстве, помогающие решить эту проблему.

При поиске в пиринговых сетях тема полноты поиска отодвигается на второй план, главная же задача – быстрое и эффективное нахождение наиболее релевантных откликов на запрос, передаваемый от рабочей станции всей сети. В частности, актуальна проблема – уменьшение сетевого трафика, порождаемого запросом (например, пересылки запроса по многочисленным рабочим станциям), и в то же время получение наилучших характеристик выдаваемых документов, т.е. получение качественного результата.

Приемлемое качество поиска в пиринговых сетях на сегодняшний день обеспечивают лишь специализированные, централизованно наполняемые, поисковые веб-сайты, естественно, работающие по протоколу HTTP. Например, для файлообменной сети Bittorrent одним из таких поисковых серверов является zagruzi.me (рис. 25).

Рис. 25 – Zagruzi.me – поиск торрентов в сети Bittorrent

Как и для файлообменных сетей, для этих серверов особо актуальными и критичными являются проблемы качества и достоверности предоставляемого контента, фальсификация файлов и распространение фальшивых

ресурсов, вирусов, «троянских коней», возможность фальсификации ID рабочих станций.

2.6. Визуализация результатов поиска

Пользователь будет считать систему «своей», если он в значительной степени может управлять результатами ее работы, вводя управляющие воздействия и получая соответствующие отклики. В случае поисковых систем воздействие пользователя – это запрос и набор дополнительных параметров. Можно констатировать, что возможности эффективного управления результатами поиска в общем случае находятся в обратной зависимости от объема индекса системы.

Например, ни одна из известных авторам систем не в состоянии сегодня ранжировать найденные документы по длине, как это делала некогда система Alltheweb в то время, когда ее индекс не превышал миллиарда документов. Система «Яндекс» позволяет ранжировать релевантные документы по датам, чего не способна делать Google, индекс которой на порядок больший. А именно последняя возможность является принципиально важной для аналитической работы – при постоянном отслеживании публикаций по одному и тому же запросу. Для пользователя существенно наличие интуитивного информационно-поискового языка, удобная навигация в результатах поиска, ранжирование результатов, получение автоматически формируемых понятных кратких описаний документов (сниппетов).

Интерфейсы режимов «расширенного поиска» большинства поисковых систем сегодня, с одной стороны, перегружены, а с другой – не позволяют своими средствами в полной мере выразить информационные потребности. Очень часто пользователь буквально тонет в результатах первичного поиска, но, переходя в режим «расширенного», заполняя все прилагаемые поля, приходит к отсутствию результатов (при том что основная проблема веба, как информационной среды, — это избыточность информации). В связи с этим в последнее время

получили распространение более гибкие визуальные интерфейсы уточнения запросов, чаще всего реализуемые путем «квазиинтеллектуальной» группировки (кластеризации) результатов первичного поиска. Появилось множество подходов, общее у которых – попытка предоставить результаты поиска и соответствующие им кластеры в удобном для пользователей виде.

Например, российская метапоисковая система Nigma.ru по словосочетанию «пиринговая сеть» представляет фильтры (термины для уточнения запроса), соответствующие автоматически определенным терминам (тэгам) «файлообменная сеть», «что такое пиринговая сеть», «P2P», «peer to peer» и др., что вполне соответствует природе данных сетей.

Другая российская поисковая система Qintura (http://quintura.ru/) обладает «интеллектуальным» интерфейсом, который обеспечивает получение автоматических подсказок по введенному запросу, помогает динамично управлять процессом поиска (рис. 26). По запросу из одного слова система Quintura предъявляет возможные фразы и словосочетания, которыми при необходимости расширяется первичный поисковый запрос.

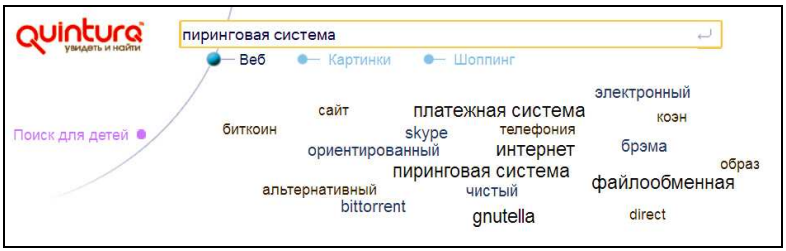


Рис. 26 – Облако тегов в системе Quintera по запросу «пиринговая система»

Сервис www.iboogie.tv – это метапоисковая система с кластеризацией, позволяющая проводить поиск почти на 50-ти языках. При поиске с помощью этой метапоисковой системы отображается автоматически

формируемый многоуровневый список тем и поисковых терминов (рис. 27).

Одним из самых динамичных направлений интернет-бизнеса сегодня можно считать оптимизацию веб-ресурсов под поисковые системы. Компания TouchGraph, в частности, реализовала интерфейс для визуализации связей между близкими (в соответствии с представлениями, заложенными в поисковую систему Google) по тематике сайтами — TouchGraph SEO Browser. Самый популярный инструмент компании TouchGraph SEO Browser (http://www.touchgraph.com/seo), представляет собой Java-апплет для визуализации тематического подобия веб-сайтов. Это весьма полезный инструмент при поиске сайтов, связанных с исходным общей тематикой. В интерфейсе TouchGraph SEO Browser можно увидеть все сайты, связанные ссылками с исходным заданным сайтом (рис. 28), при этом пользователь может задавать глубину связей и отображать взаимосвязи различных веб-сайтов.

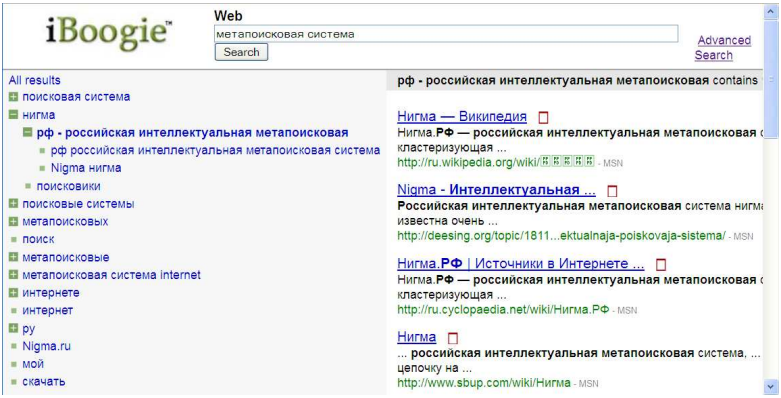


Рис. 27 – Иерархия тем по запросу «метапоисковая система»

Системы визуального поиска фокусируются на психологических аспектах человеческого восприятия, ориентируясь на методики, которые используют люди в

процессе поиска. Именно поэтому визуальные поисковые системы имеют все шансы потеснить на информационном рынке таких гигантов веб-поиска, как Google и Yahoo!, используя базы данных последних.

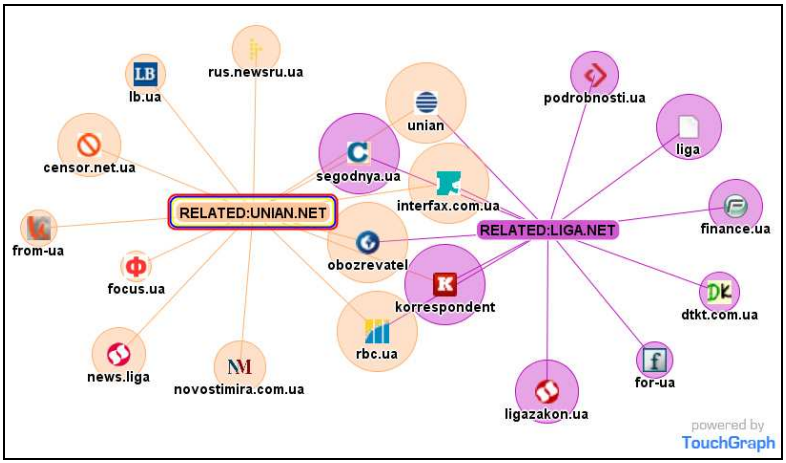


Рис. 28 – Интерфейс системы TouchGraph SEO Browser

С другой стороны, может быть, более очевидный ход событий заключается в переходе ведущих сетевых ИПС в область визуального поиска? Ведь для пользователей визуально организованные результаты поиска выглядят гораздо привлекательнее и понятнее, чем списки гиперссылок и сниппетов, формируемых традиционными поисковыми системами.

3. СОДЕРЖАТЕЛЬНЫЙ АНАЛИЗ ИНФОРМАЦИОННЫХ ПОТОКОВ

3.1. Семантическая обработка информации

Главная проблема современных коммуникаций – извлечение действительно ценных сведений из информационных потоков, или, другими словами, получение знаний из информации. В настоящее время сложилось понимание того, что для решения проблемы аналитической обработки документальной текстовой информации в глобальной сетевой среде больше подходят технологии, порожденные когда-то таким направлением, как контент-анализ, который получил сегодня название Data Mining и Text Mining [Барсегян, 2007], [Mostafa, 2013].

Технологии глубинной разработки текста исторически предшествовало создание технологии глубинной добычи данных (Data Mining), методология и подходы которой широко используются в методах Text Mining **[**Самойленко, 2001**],** **[**Чубукова, 2006]. Актуальность технологии Data Mining состоит в том, что она позволяет дополнительно в «сырых данных» обнаружить практически полезные знания, необходимые для принятия решений в различных сферах трудовой деятельности.

Data Mining – это процесс выделения из данных неявной и неструктурированной информации и представление ее в виде, пригодном для реализации. Data Mining – это процесс, цель которого – найти новые значимые корреляции, образцы и тенденции в результате просеивания большого объема хранимых данных с использованием методик распознавания образов плюс применение статистических и математических методов. Data Mining дословно переводится как «добыча» или «раскопка данных». Однако нередко используют и другие термины: «выявление знаний в базах данных» или «интеллектуальный анализ данных», которые можно считать синонимами Data Mining. Идеология Data

Mining появилась на стыке прикладной статистики, искусственного интеллекта и баз данных. Фактически рождению нового направления при анализе данных способствовало появление мощных компьютеров и совершенствование технологий записи и хранения данных.

Системы добычи данных в большей степени ориентированы на практическое применение полученных результатов, чем на выяснение природы явления, которое исследуется. При использовании технологии Data Mining исследователя-аналитика не очень интересует конкретный вид зависимостей между переменными решаемой задачи. Основное внимание уделяется поиску решений, на основе которых можно было бы строить достоверные прогнозы. В процедуре добычи данных преобладает содержательный подход к изучению знаний, а критерием качества методов, которые применяются, является их практическая реализация. Новое направление в обработке текстовой информации – «глубинная разработка текстов» (Text Mining) – это алгоритмическое выявление ранее неизвестных связей и корреляций в уже имеющихся текстовых данных. Задача Text Mining – отобрать ключевую и наиболее значимую информацию для пользователя.

Оформившись в середине 90-х годов XX-го века как направление анализа неструктурированных текстов, технология Text Mining сразу же взяла на вооружение методы классической добычи данных, например, таких как классификация или кластеризация. Разработанные на основе статистического и лингвистического анализа, а также методов искусственного интеллекта технологии Text Mining предназначены для проведения смыслового анализа.

В Text Mining появились и дополнительные возможности, такие как автоматическое реферирование текстов и выявление феноменов, т.е. понятий и фактов. Важная компонента технологии Text Mining связана с

извлечением из текста характерных элементов или признаков, которые могут использоваться в качестве ключевых слов, метаданных, аннотаций. Еще одна задача Text Mining – отнесение документов к некоторым категориям из заданной схемы их систематизации. Кроме того, Text Mining – это новый вид поиска, который в отличие от традиционных подходов не только находит списки документов, формально релевантных запросам, но и помогает в понимании смысла текстов. Таким образом, пользователю будет незачем самому "просеивать" огромное количество неструктурированной информации. Text Mining – это алгоритмическое выявление прежде не известных связей в уже имеющихся данных. Применяя технологию Text Mining, пользователи могут получать новую ценную информацию – знания.

Следует подчеркнуть, что технологии Text Mining свойственна абсолютная объективность – в ней отсутствует субъективизм, свойственный человеку – аналитику. Технологии Data Mining и Text Mining являются новой тенденцией в развитии средств и методов обработки данных. Они помогают найти скрытые закономерности и отношения в данных, исследуемых для того, чтобы можно было принять более обоснованные решения. Сфера их применения ничем не ограничена – они везде, где есть какие-либо данные. Но, в первую очередь, в применении новых технологий на сегодняшний день наиболее заинтересованы коммерческие предприятия, которые проектируют и внедряют проекты на основе информационных хранилищ данных. Новые технологии представляют большую ценность для руководителей и аналитиков в их повседневной деятельности. Опыт многих таких предприятий показывает, что отдача от использования технологий Data Mining и глубинного анализа текстов Text Mining может достигать 1000% [Самойленко, 2001].

Технология, предшествовавшая глубинному анализу текстов – контент-анализ, начиналась как количественно-ориентированный метод анализа текстов для изучения массовых коммуникаций. Впервые он был

применен в 1910 году социологом Максом Вебером (Max Weber) для оценки освещения печатными изданиями политических акций в Германии. Американский исследователь средств коммуникации Гарольд Лассвелл (Harold Lasswell) в 30-40-е годы использовал подобную методику для изучения содержания пропагандистских сообщений военного времени.

С появлением средств автоматизации, текстов в электронном виде, начиная с 60-х годов прошлого века, начальное развитие получил контент-анализ информации больших объемов – баз данных и интерактивных медиа-источников. Традиционное политическое использование современных технологий контент-анализа было дополнено неограниченным перечнем рубрик, охватывающих производственную и социальную сферы, бизнес и финансы, культуру и науку. Этот процесс, в свою очередь, сопровождался большим количеством разнородных программных систем **[**Сорока, 1998**].**

Понятие контент-анализа, которое берет свое начало в психологии и социологии, сегодня пока еще не имеет однозначного определения. Это порождает ряд проблем, важнейшая из которых заключается в том, что программные системы, построенные на основе различных подходов к контент-анализу, в общем случае несовместимы. Приведем лишь некоторые определения контент-анализа:

Контент-анализ – методика объективного качественного и систематического изучения содержания средств коммуникации (Д. Джерри, Дж. Джерри).

Контент-анализ – систематическая числовая обработка, оценка и интерпретация формы и содержания информационного источника (Д. Мангейм, Р. Дело).

Контент-анализ – качественно-количественный метод изучения документов, который характеризуется объективностью выводов и строгостью процедуры и

представляет собой квантификационную обработку текста с последующей интерпретацией результатов (В. Иванов).

Контент-анализ состоит из поиска в тексте определенных содержательных понятий (единиц анализа), выявления частоты их появления и соотношения с содержанием всего документа (Б. Краснов).

Контент-анализ – исследовательская техника для получения результатов путем анализа содержания текста о состоянии и свойствах социальной действительности (Э. Таршис).

Большинство из приведенных выше определений конструктивные, то есть процедурные. Через различные подходы они порождают различные алгоритмы, которые порой противоречат друг другу. Существующие различные подходы к пониманию контент-анализа подвергаются вполне оправданной критике. Наибольшие сомнения вызывает игнорирование роли контекста. Тем не менее, несмотря на многообразие трактовок контент-анализа, большое прикладное значение методологии все-таки позволяет избежать многих противоречий. Объединение средств и методов, их естественный отбор путем многократной оценки полученных результатов открывают возможность выделения и подтверждения знаний, а также фактической силы и полезности этого инструментария.

Диапазон методов и процедур, которые касаются самого процесса контент-анализа, очень широк. Но наиболее важным при подготовке исследования, как показал опыт, является выполнение следующих действий:

- описание проблемной ситуации, поиск цели исследования;
- точное определение объекта и предмета исследования;
- предварительный анализ объекта;

- содержательное уточнение и эмпирическая интерпретация понятий;
- описание процедур регистрации свойств и явлений;
- определение общего плана исследования;
- определение типа выборки, круга источников и т.п.

Интересной особенностью контент-анализа является и то, что эту методологию до последнего времени связывали с определенной сферой человеческой деятельности (политикой и социологией). Тем не менее, на сегодня контент-анализ все шире применяется во многих сферах политической и экономической жизни, способствует большему прикладному значению использованных в методологии контент-анализа философских категорий, социологии и лингвистики.

Контент-анализ в рамках исследования информационных потоков – новое направление, которое предусматривает анализ массива текстовых документов – результатов мониторинга информационного пространства.

Общепризнанным является разделение методологии контент-анализа на две ветви: качественную и количественную. Основа количественного контент-анализа – частота появления в документах определенных характеристик содержания. Метод качественного контент-анализа базируется на самом факте присутствия или отсутствия в тексте одной или нескольких характеристик содержания.

Метод качественного контент-анализа основан на том, что в любой фазе количественного контент-анализа для оценок результатов может быть привлечен эксперт. Таким образом, этот метод призван обеспечить эксперта-аналитика необходимыми средствами для выводов и дополнительных результатов. Эксперт с помощью таких средств может выявить определенные

свойства части информации и проверить их относительно общего текстового потока, а общие свойства текстового потока распространить на его конкретную тематическую часть.

Процесс метода качественного контент-анализа состоит из трех основных стадий. Первая – сведение большого количества текстовой информации к конечному числу интегрированных блоков текста – единиц содержания, которые кодируются для дальнейшей обработки этих блоков. Основными единицами содержания являются категории, последовательности и темы. Вторая стадия качественного контент-анализа – реконструкция субъективных составляющих текстового потока – системы значений, мыслей, взглядов и доказательств каждого источника текста. Третья стадия – формирование выводов и обобщений.

Метод количественного контент-анализа, в свою очередь, как правило, состоит из трех основных этапов. На первом этапе выделяются единицы анализа и переводятся в форму, приемлемую для обработки (сегодня – в электронный вид). Второй этап заключается в подсчете частот единиц анализа с применением разнообразного математического аппарата для выявления взаимосвязей между ними. Суть третьего этапа заключается в интерпретации полученных результатов. При этом, без привлечения искусственного интеллекта, объемных семантических формализаторов, даже экспертов как таковых, с использованием только математических методов могут быть получены содержательные, семантически наполненные результаты.

Таким образом, в простейшем виде идею контент-мониторинга можно сформулировать как постоянное выполнение узко очерченного своими задачами контент-анализа непрерывных информационных потоков. Именно непрерывное воспроизведение во времени процесса обработки входных данных является характерной особенностью контент-мониторинга.

Собственно контент-анализ выступает здесь как составляющая, а контент-мониторинг имеет собственную проблематику и собственные пути решения прикладных задач.

Основные элементы Text Mining. Обычно выделяют четыре основных вида приложений технологии Text Mining.

1. Классификация текста, в которой используются статистические корреляции для построения правил размещения документов в определенные категории.

2. Кластеризация, базирующаяся на признаках документов. Используются лингвистические и математические методы без применения определенных категорий.

3. Построение семантической сети или анализ связей, определяющих появление дескрипторов (ключевых слов и словосочетаний) в документе для обеспечения поиска и навигации.

4. Извлечение фактов из текста с целью улучшения классификации, поиска и кластеризации.

Наиболее часто решаемая в Text Mining задача – это классификация, т.е. отнесение объектов базы данных к заранее определенным категориям. Фактически задача классификации – это вариант классической задачи распознавания, когда система по обучающей выборке относит новый объект к той или иной категории.

Особенность же системы Text Mining заключается лишь в том, что количество таких объектов и их атрибутов может быть очень большим. Поэтому должны быть предусмотрены интеллектуальные механизмы оптимизации процесса классификации.

В существующих сегодня системах классификация применяется, например, для решения таких задач, как группировка документов, размещение документов в определенные папки, сортировка сообщений

электронной почты, избирательное распространение новостей подписчикам и др.. Другая задача, основанная на кластеризации, состоит в выделении компактных подгрупп объектов с близкими свойствами. Система должна самостоятельно найти признаки и разделить объекты по подгруппам. Решение этой задачи, как правило, предшествует задаче классификации, поскольку позволяет определить группы объектов. В процессе кластеризации строится базис ссылок от документа к документу, основанный на весах и совместном применении обусловленных ключевых слов.

Сегодня кластеризация широко применяется при реферировании больших документальных массивов или определении взаимосвязанных групп документов, а также для упрощения процесса просмотра при поиске необходимой информации, для нахождения уникальных документов из коллекции, для выявления дубликатов или очень близких по содержанию документов.

К числу задач, которые можно решать средствами технологии Text Mining, можно отнести, например, прогнозирование, которое состоит в том, чтобы предугадывать по значениям одних признаков значения других.

Еще одна задача – нахождение исключений, т.е. поиск объектов, которые по своим характеристикам сильно выделяются из общей массы. Для этого сначала выясняются средние параметры объектов, а затем исследуются те объекты, параметры которых наиболее сильно отличаются от средних значений. Как правило, поиск исключений проводится после классификации или кластеризации – для того чтобы выяснить, насколько были точны последние.

Несколько особняком от кластеризации находится задача поиска связанных признаков (полей, понятий) отдельных документов. От прогнозирования эта задача отличается тем, что заранее неизвестно, по каким именно признакам реализуется взаимосвязь. Цель именно в том и заключается, чтобы найти связь между отдельными признаками. Эта задача подобна

кластеризации, но выполняется не по множеству документов, а по множеству признаков, присущих документу.

И, наконец, для обработки и интерпретации результатов Text Mining большое значение приобретает визуализация данных, что подразумевает обработку структурированных числовых данных. Однако визуализация также является ключевым звеном при представлении данных неструктурированных текстовых документов. В частности, современные системы класса Text Mining могут осуществлять анализ больших массивов документов и формировать предметные указатели понятий и тем, освещенных в этих документах.

Визуализация обычно используется как средство представления контента всего массива документов, а также для реализации навигационного механизма, который может применяться при исследовании документов и их классов.

3.2. Классификация информации

Все традиционные модели информационного поиска имеют общий недостаток по представлению данных, который характеризуется большой размерностью векторного пространства (векторная модель) и множества (булева и вероятностная модель). Для обеспечения эффективной работы ИПС необходимо сгруппировать как подмножества термов, так и тематически подобные документы. Только в этом случае может быть обеспечена обработка информационных массивов в режиме реального времени. В этом случае используются два основных приема группировки – *классификация и кластеризация.*

Классификация и кластеризация представляют собой два противоположных полюса применительно к участию пользователя в процессе группировки документов в информационной сети. Осуществление на практике возможности автоматической группировки тематически близких документов позволяет выстроить

тематические каталоги. Механизм классификации обычно устанавливается на отобранных документах только после завершения стадии автоматического обнаружения сгруппированных данных (кластеров).

Классификация текста, в которой используются статистические корреляции для построения правил размещения документов в определенные категории. Под классификацией текстов (Text Categorization, TC) понимается распределение текстовых документов по заранее определенным категориям (в противоположность кластеризации, где множество категорий заранее неизвестны).

Кластеризация, базирующаяся на признаках документов, использует лингвистические и математические методы без использования определенных категорий.

Методы классификации текстов лежат на стыке двух областей – машинного обучения (machine learning, ML) и информационного поиска (information retrieval, IR).

Соответственно, автоматическая классификация может осуществляться: на основе заранее заданной схемы классификации и уже имеющегося множества классифицированных документов; полностью автоматизированно.

Задача классификации текстов заключается в определении принадлежности текста, который рассматривается, одному или нескольким классам. Классификация может определяться общей тематикой текстов, наличием определенных дескрипторов или выполнением определенных условий, иногда довольно сложных.

Для каждого класса эксперты отбирают текстовые массивы (наборы типовых документов), используемых системой классификации в режиме обучения. После того, как обучение закончено, система с помощью специальных алгоритмов сможет распределять

входящие потоки текстовой информации по классам.

Классификацию можно рассматривать как задачу распознавания образов, при таком подходе для каждого объекта выделяются наборы признаков. В случае текстов признаками являются слова и взаимозависимые наборы слов – термы, которые содержатся в текстах. Для формирования набора признаков для каждого документа используются лингвистические и статистические методы. Признаки группируются в специальную таблицу – информационную матрицу. Каждая строка информационной матрицы соответствует одному из классов, каждый элемент строки – одному из признаков; численное значение этого элемента определяется в процессе обучения системы классификации. Когда обучение завершается, принадлежность нового текста к одному из классов устанавливается путем анализа признаков этого текста с учетом соответствующих весовых значений. Существующие алгоритмы позволяют осуществлять классификацию с достаточно высокой точностью, однако результаты достигаются за счет больших размеров информационной матрицы, которая определяется общим числом дескрипторов – термов.

Автоматическая классификация может применяться в таких процедурах информационного поиска, как:

- фильтрация (избирательный отбор) информации;
- формирование тематических каталогов;
- поиск по классам;
- реализация обратной связи по релевантности путем классификации результатов поиска и выбора пользователем релевантных классов;
- расширение запросов за счет термов, характеризующих тематику класса;
- снятие омонимии (т.е. учет тех случаев, когда

одно и то же слово может иметь разный смысл);

- автоматическое реферирование.

3.2.1. Формальное описание классификации

Пусть $D=\{d_1, ..., d_{|D|}\}$ - множество объектов (узлов сети или, например, их содержательных элементов – документов), $C=\{c_1, ..., c_{|C|}\}$ – множество категорий, Ф – целевая функция, которая по паре $< d_i, c_j >$ определяет, относится ли документ d_i к категории c_j (1 или True) или нет (0 или False). Задача классификации состоит в построении функции Ф', максимально близкой к Ф.

Методы машинного обучения, которые применяются для классификации, предусматривают наличие коллекции заранее классифицированных экспертами объектов, т.е. таких, для которых уже точно известно значение целевой функции. Для того чтобы после построения классификатора можно было оценить его эффективность, эта коллекция разбивается на две части, не обязательно равного размера:

1. Обучающая (training-and-validation, TV) коллекция. Классификатор Ф' строится на основе характеристик этих объектов.

2. Тестовая (test, Te) коллекция. На ней проверяется качество классификации. Объекты из Te не должны использоваться в процессе построения классификатора.

Рассматриваемая классификация называется четкой бинарной, т.е. подразумевается, что существуют только две категории, которые не пересекаются. К такой классификации сводится много задач, например, классификация по множеству категорий $C=\{c_1, ..., c_{|C|}\}$

разбивается на $|C|$ бинарных классификаций по множествам $\{c_i, \overline{c}_i\}$.

Часто используется ранжирование, при котором множество значений целевой функции – это отрезок [0, 1]. Объект при ранжировании может относиться не только к одной, а сразу к нескольким категориям с разной степенью принадлежности, т.е. категории могут пересекаться между собой.

3.2.2. Ранжирование и четкая классификация

Предположим, что для каждой категории c_i построена фунция CSV_i. Рассмотрим задачу, заключающуюся в том, чтобы от функции ранжирования перейти к точной классификации. Наиболее простой способ – для каждой категории c_i выбрать предельное значение (порог) τ_i. Если $CSV_i(d) > \tau_i$, то документ d соответствует категории c_i. Другой подход: для каждого документа d выбирать k ближайших категорий, т.е. k категорий, на которых $CSV_i(d)$ принимают наибольшие значения.

Выбирать пороговое значение можно несколькими способами:

- Пропорциональный метод. Обучающая коллекция разбивается на две части. Для каждой категории c_i на одной части учебной коллекции вычисляется, какая часть документов ей принадлежит. Пороговые значения выбираются так, чтобы на другой части учебной коллекции количество оставшихся документов, отнесенных c_i, было таким же.

- Метод k ближайших категорий. Каждый документ d_i считается принадлежащим к k ближайшим категориям и соответственно этому выбирается пороговое значение.

3.2.3. Мера близости объекта и категории

Пусть каждой категории C_i соответствует вектор $\boldsymbol{C_i} = (c_{i1}, ..., c_{iN})$, где N – размерность пространства термов. В качестве правила классификатора может использоваться скалярное произведение:

$$CSV_i(d) = \boldsymbol{d} \cdot \boldsymbol{C}_i = \sum_{j=1}^{N} c_{ij} d_j .$$

Нормализация проводится обычно таким образом, чтобы итоговая формула для $CSV_i(d)$ представляла собой скалярное произведение – косинус угла между вектором категории $\boldsymbol{C_i}$ и вектором из весовых значений термов, входящих в документ d – $\boldsymbol{d} = (d_1, ..., d_N)$:

$$CSV_i(d) = \frac{\boldsymbol{d} \cdot \boldsymbol{C_i}}{|\boldsymbol{d}| \cdot |\boldsymbol{C_i}|}.$$

Координаты вектора $\boldsymbol{C_i}$ определяются в ходе обучения, которое проводится по каждой категории независимо от других.

3.2.4. Метод Rocchio

Некоторые классификаторы используют так называемый профайл для определения категории. Профайл – это список взвешенных термов, присутствие или отсутствие которых позволяет наиболее точно отличать конкретную категорию от других категорий. К таким методам классификации относится и метод Rocchio, который относится к линейным классификаторам, в которых каждый документ представляется в виде вектора весовых значений термов. Профайл категории i будем рассматривать как

вектор $\boldsymbol{C_i} = (c_{1i}, ..., c_{Ni})$ (N – количество термов в словаре), значения элементов которого c_{ki} в рамках метода Rocchio рассчитывается по формуле:

$$c_{ki} = \frac{\alpha}{|POS_i|} \cdot \sum_{d_j \in POS_i} w_{kj} - \frac{\beta}{|NEG_i|} \cdot \sum_{d_j \in NEG_i} w_{kj}, \qquad (2.4.3)$$

где w_{kj} – это вес терма t_k в документе d_j (рассчитанный, например, по принципу *TF IDF*), $POS_i = \{d_j \mid \Phi(d_j, c_i) = 1\}$ и $NEG_i = \{d_j \mid \Phi(d_j, c_i) = 0\}$. В этой формуле, α и β – контрольные параметры, которые характеризуют значимость положительных и отрицательных примеров. Например, если $\alpha = 1$ и $\beta = 1$, то $\boldsymbol{C_i}$ будет центром масс всех документов, относящихся к соответствующей категории.

Функция $CSV_i(d)$ определяется либо как величина, обратная расстоянию от вектора из весовых значений термов, входящих в документ d, до профайла категории i – $\boldsymbol{C_i}$, либо как скалярное произведение этих векторов. Метод Rocchio дает удовлетворительные результаты в случае, когда документы из одной категории близки друг к другу по расстоянию.

3.2.5. Метод линейной регрессии

Регрессионный анализ используется в случае, когда признаки категорий могут быть выражены количественно в виде некоторой комбинации векторов весовых значений термов, входящих в документы из обучающей коллекции. Полученная комбинация может использоваться для определения категории, к которой будет относиться новый документ. В простейшем случае для решения этой задачи используются стандартные статистические методы, такие как линейная регрессия.

Метод регрессии является вариантом линейной классификации, обучаемой сразу на всей коллекции. При применении регрессионного анализа к

классификации текстов рассматривается множество термов (F) и множество категорий (C). В этом случае обучающей коллекции документов ставятся в соответствие две матрицы:

- матрица документов $\boldsymbol{D}$ в обучающей коллекции, в которой каждая строка – это документ, а столбец – терм, количество строк N – количество документов в обучающей коллекции;
- матрица ответов $\boldsymbol{O} = \left\| o_{i,j} \right\|$, в которой строка i соответствует документу ($i = 1, \ldots, N$), столбец j – категории ($j = 1, \ldots, K$), а $O_{i,j}$ – значению $CSV_j(d_i)$.

Метод регрессии базируется на алгоритме нахождения матрицы правил $\boldsymbol{M}$, которая минимизирует значение нормы матрицы $\left\| \boldsymbol{MD - O} \right\|_F$, то есть:

$$M = \arg \min_M \left\| \boldsymbol{MD - O} \right\|_F .$$

Напомним, что в линейной алгебре под нормой матрицы понимается функция, которая ставит в соответствие матрице числовую характеристику. Норма матрицы отражает порядок величины матричных элементов. В данном случае рекомендуется использовать норму Фробениуса $\left\| \cdot \right\|_F$, равную корню квадратному из суммы квадратов всех элементов соответствующей матрицы:

$$\left\| A \right\|_F = \sqrt{\sum_{i,j} a_{ij}^{\ 2}}.$$

Элемент m_{ij} искомой матрицы $\boldsymbol{M}$ будет отражать степень принадлежности i–го терма j–й категории.

3.2.6. Байесовская логистическая регрессия

В модели байесовской логистической регрессии рассматривается условная вероятность принадлежности документа D классу C: $p(C|D)$.

Предполагается, что документ определяется термами, входящими в него, т.е. в рамках данной модели документ – это вектор $D=(w_1, ..., w_N)$, где w_i – вес терма i, а N – размер словаря.

Модель байесовской логистической регрессии задается формулой:

$$p(C|D)=\varphi(\beta \cdot D)=\varphi(\sum_{i=1}^{N}\beta_i \cdot w_i),$$

где $C \in \{0,1\}$, $\beta=\{\beta_1, ..., \beta_N\}$ - вектор параметров модели, а φ – логистическая функция, в качестве которой рекомендуется использовать:

$$\varphi(x)=\frac{1}{1+\exp(-x)}.$$

Основная идея подхода состоит в том, чтобы использовать предшествующее распределение вектора параметров β, в котором каждое конкретное значение β_i с большой вероятностью может принимать значение, близкое к 0. При реальных расчетах принимаются гипотезы о Гауссовском или Лапласовом распределении значений β_i, а также то, что все величины β_i взаимно независимы.

3.2.7. Метод опорных векторов

Метод опорных векторов (Support Vector Mashine, SVM), предложенный В.Н. Вапником [Vapnik, 1998], относится к группе граничных методов классификации. Он определяет принадлежность объектов к классам с помощью границ областей.

Рассматривается бинарная классификация, т.е. только по двум категориям c и $\overline{c}$ (принимается во внимание то, что этот подход может быть расширен на любое конечное число категорий). Кроме того предполагается, что каждый объект классификации является вектором в N-мерном пространстве. Каждая координата вектора – это некоторый признак, количественно тем больший, чем больше этот признак выражен в данном объекте.

Предполагается, что существует обучающая коллекция – это множество векторов $\{x_1, ..., x_n\} \in R^N$ и чисел $\{y_1, ..., y_n\} \in \{-1, 1\}$. Число y_i равно 1 в случае принадлежности соответствующего вектора x_i категории c, и -1 – в противном случае. Как было показано выше, линейный классификатор – это один из простейших способов решения задачи классификации. В этом случае ищется прямая (гиперплоскость в N-мерном пространстве), отделяющая все точки одного класса от точек другого класса. Если удается найти такую прямую, то задача классификации сводится к определению взаимного расположения точки и прямой: если новая точка лежит с одной стороны прямой (гиперплоскости), то она принадлежит классу c, если с другой стороны – классу $\overline{c}$.

Формализуем эту классификацию: необходимо найти вектор $\boldsymbol{w}$ такой, что для некоторого предельного значения b и новой точки x_i выполняется:

$$y_i = \begin{cases} +1, \ \textit{если } \boldsymbol{w} \cdot \boldsymbol{x_i} \geq b, \\ -1, \ \textit{если } \boldsymbol{w} \cdot \boldsymbol{x_i} < \boldsymbol{b}, \end{cases}$$

где $\boldsymbol{w} \cdot \boldsymbol{x}_i$ – скалярное произведение векторов $\boldsymbol{w}$ и $\boldsymbol{x}_i$:

$$\boldsymbol{w} \cdot \boldsymbol{x}_i = \sum_{j=1}^{N} w_j x_{i,j}.$$

Уравнение $\boldsymbol{w} \cdot \boldsymbol{x}_i = b$ описывает гиперплоскость, которая разделяет классы. То есть, если скалярное произведение вектора w на x_i не меньше значения b, то новая точка принадлежит первому классу, если меньше – второму. Известно, что вектор $\boldsymbol{w}$ перпендикулярен искомой разделяющей прямой, а значение b зависит от кратчайшего расстояния между разделяющей прямой и началом координат. Очевидно, если существует одна разделяющая прямая, то она не единственная. Возникает вопрос, какая из прямых разделяет классы лучше всего?

Метод SVM базируется на таком постулате: наилучшая разделяющая прямая – это та, которая максимально далеко отстоит от ближайших до нее точек обоих классов. То есть задача метода SVM состоит в том, чтобы найти такие вектор w и число b, чтобы для некоторого $\varepsilon > 0$ (половины ширины разделяющей поверхности) выполнялось:

$$\begin{cases} \boldsymbol{w} \cdot \boldsymbol{x}_i \geq b + \varepsilon \;\Rightarrow\; y_i = +1, \\ \boldsymbol{w} \cdot \boldsymbol{x}_i \leq b - \varepsilon \;\Rightarrow\; y_i = -1. \end{cases}$$

Умножим после этого обе части неравенства на $1/\varepsilon$ и, не ограничивая общности, выберем ε равным единице. Таким образом, для всех векторов x_i из обучающей коллекции будет справедливо:

$$\begin{cases} w \cdot x_i - b \geq +1, \text{ если } y_i = +1, \\ w \cdot x_i - b \leq -1, \text{ если } y_i = -1. \end{cases}$$

Условие $-1 < w \cdot x_i - b < 1$ задает полосу, которая разделяет классы. Границами полосы являются две параллельные гиперплоскости с направляющим вектором w. Точки, ближайшие к разделяющей гиперплоскости, расположены точно на границах полосы.

Чем шире полоса, тем увереннее можно классифицировать документы, соответственно, в методе SVM предполагается, что самая широкая полоса является наилучшей.

Сформулируем условия задачи оптимальной разделяющей полосы, определяемой неравенством: $y_i(w \cdot x_i - b) \geq 1$ (так переписывается система уравнений, исходя из того, что $y_i \in \{-1, 1\}$). Ни одна из точек обучающей выборки не может лежать внутри этой разделяющей полосы. При этих ограничениях x_i и y_i – постоянные, как элементы обучающей коллекции, а w и b – переменные.

Из геометрических соображений известно, что ширина разделяющей полосы равна $2/\|w\|$. Поэтому необходимо найти такие значения w и b, чтобы выполнялись приведенные линейные ограничения, и при этом как можно меньше была норма вектора w, то есть необходимо минимизировать:

$$\|w\|^2 = w \cdot w.$$

Это известная задача квадратичной оптимизации при линейных ограничениях.

Метод классификации разделяющей полосой имеет два недостатка:

- при поиске разделяющей полосы важное значение имеют только пограничные точки;
- во многих случаях найти оптимальную разделяющую полосу невозможно.

Для улучшения метода применяется идея расширенного пространства, для чего:

- Выбирается отображение $\phi(x)$ векторов $\boldsymbol{x}$ в новое, расширенное пространство.

- Автоматически применяется новая функция скалярного произведения, которая применяется при решении задачи квадратичного программирования, так называемая функция ядра (kernel function): $K(x,y)=\phi(x)\cdot\phi(y)$. На практике обычно выбирают не отображение $\phi(x)$, а сразу функцию $K(x,y)$, которая могла бы быть скалярным произведением при некотором отображении $\phi(x)$. Функция ядра – главный параметр настраивания машины опорных векторов.
- Находим разделяющую гиперплоскость в новом пространстве: с помощью функции $K(x,y)$ устанавливается новая матрица коэффициентов для задачи оптимизации. При этом вместо $x_i \cdot x_j$ подставляются значения $K(x_i,x_j)$, и решается новая задача оптимизации.
- Найдя $\boldsymbol{w}$ и b, получаем поверхность, которая классифицирует $w\cdot\phi(x)-b$ в новом, расширенном пространстве.

Например, в системе классификации новостного контента с применением известного пакета LibSVM (http://www.csie.ntu.edu.tw/~cjlin/libsvm) в качестве функции ядра рекомендуется использовать радиальную базисную функцию:

$$K(x,y)=\exp(-\gamma\|x-y\|^2),$$

где γ – настраиваемый параметр.

Рассмотрим наглядный пример перехода к расширенному пространству, изображенный на рис. 29. По всей видимости, круглые и квадратные фигуры не разделяются линейной полосой. Если же «изогнуть» пространство, перейдя к третьему измерению, то эти фигуры можно разделить плоскостью, которая отрезает часть поверхности с квадратными точками. Таким

образом, выгнув пространство с помощью отображения $\phi(x)$, можно найти разделяющую гиперплоскость.

К недостаткам можно отнести:

- мало параметров для настройки;
- нет четких критериев выбора ядра;
- довольно медленное обучение системы классификации.

Метод SVM обладает такими преимуществами:

- на тестах с документальными массивами превосходит другие методы;
- при выборах разных ядер позволяет эмулировать другие подходы. Например, большой класс нейронных сетей можно представить с помощью SVM с определенными ядрами;
- итоговое правило выбирается не с помощью некоторых эвристик, а путем оптимизации некоторой целевой функции.

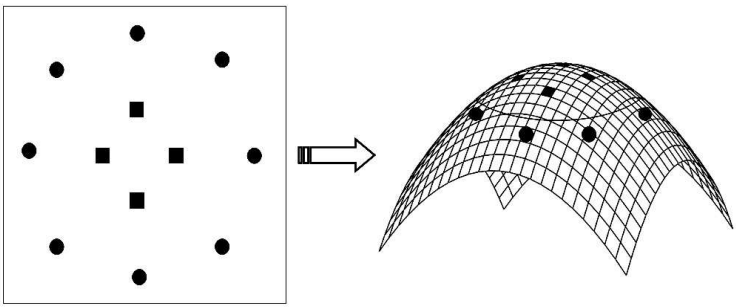


Рис. 29 – Пример перехода к расширенному пространству

3.3. Кластерный анализ

Существует четкое различие между

классификацией и кластеризацией документов. Классификация – это отнесение каждого документа в определенный класс с заранее известными параметрами, полученными на этапе обучения. Число классов строго ограничено. Кластеризация – разбиение множества документов на кластеры – подмножества, параметры которых заранее неизвестны.

Задача кластеризации – выделение компактных подгрупп объектов с близкими свойствами. Система должна самостоятельно найти признаки и разделить объекты по подгруппам. Она, как правило, предшествует задаче классификации, поскольку позволяет определить группы объектов. Различают два основных типа кластеризации – иерархическую и бинарную. Иерархическая кластеризация состоит в построении дерева кластеров, в каждом из которых размещается небольшая группа документов. В процессе кластеризации строится базис ссылок от документа к документу, основанный на весе и общем употреблении обусловленных ключевых слов. Кластеризация сегодня применяется при реферировании больших документальных массивов, определении взаимосвязанных групп документов, упрощении процесса пересмотра при поиске необходимой информации, нахождении уникальных документов из коллекции, выявлении дубликатов или очень близких по содержанию документов.

Процесс кластеризации – это разбиение множества документов на кластеры, которые являются, по сути, подмножествами с содержательными параметрами, которые заранее известны. Количество кластеров может быть произвольно или фиксировано. Основная идея современных методов кластеризации – это снижение размерности пространства признаков, по которым происходит классификация документов. Задачей кластеризации документов является автоматическое выявление групп подобных документов, исходя из их семантики. Цель всех методов кластеризации заключается в том, чтобы сходство документов, находящихся в конкретном кластере, было

бы максимально близко по своей семантике.

Поэтому методы кластерного анализа базируются на таких определениях кластера, как множества документов, значение семантической близости между любыми двумя элементами которых не меньше определенного порога или значение близости между любым документом множества и центром кластера также не меньше определенного порога.

При использовании численных методов кластерного анализа определения близости используются такие основные метрики:

Евклидово расстояние:

$$D(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sqrt{\sum_{k=1}^{N} (x_{ik} - x_{jk})^2},$$

которое является частным случаем метрики Минковского при $p = 2$:

$$D_p(\boldsymbol{x}_i, \boldsymbol{x}_j) = \left(\sum_{k=1}^{N} (x_{ik} - x_{jk})^p \right)^{\frac{1}{p}}.$$

Для группирования документов, представленных в виде векторов весовых значений входящих в них термов, часто используется метрика, базирующаяся на скалярном произведении весовых векторов:

$$Sim(\boldsymbol{x}_i,\ \boldsymbol{x}_j) = \hat{\boldsymbol{x}}_i \cdot \hat{\boldsymbol{x}}_j = \sum_{k=1}^{N} \hat{x}_{ik} \cdot \hat{x}_{jk},$$

где $\boldsymbol{x}_i$, $\boldsymbol{x}_j$ – документы x_{ik} – элемент матрицы весовых значений термов, входящих в $\boldsymbol{x}_i$, $(i = 1, ..., N)$, $\hat{\boldsymbol{x}}_i$ – нормализированный вектор $\hat{\boldsymbol{x}}_i = \boldsymbol{x}_i / |\boldsymbol{x}_i|$.

Начальным пространством признаков обычно выбирается пространство термов, которое образуется в результате анализа большого массива документов. Для проведения такого анализа используются разные подходы – весовой, вероятностный, семантический и т. д.

В области информационного поиска кластерный анализ чаще всего применяется для решения двух задач – группирования документов в базах данных (информационных массивах) и группирования результатов поиска.

Для статических документальных массивов методы кластерного анализа в настоящее время получили большое развитие и популярность. Вместе с тем открытым остается вопрос применения этих методов к динамично изменяемым информационным потокам, которым присущи, кроме динамики, еще и большие объемы.

Методы кластерного анализа находят широкое применение в процедурах ранжирования откликов информационно-поисковых систем, при построении персонализированных папок поиска, персональных поисковых интерфейсов пользователей информационно-поисковых систем.

3.3.1. Метод k-means

Итеративный алгоритм кластерного анализа *k*-means (*k*-средних) группировки документов по фиксированному количеству кластеров заключается в следующем: случайным образом выбирается k векторов, которые определяются как центроиды (наиболее типичные представители) кластеров. Затем k кластеров $\{C_1, C_2, ..., C_k\}$ наполняются – для каждого из векторов, которые остались, некоторым образом определяется близость к центроиду соответствующего кластера.

Близость может определяться разными способами, в частности, как нормированное скалярное произведение:

$$Sim(\boldsymbol{x},\ \boldsymbol{c}^j) = \frac{\sum_{k=1}^{N} x_k c_k^j}{\left|\boldsymbol{x}\right|\left|\boldsymbol{c}^j\right|},$$

где $\boldsymbol{x}$ – документ, $\boldsymbol{c}^j$ $(j = 1,\ ...,\ k)$ – центроид кластера C_j, N – размерность пространства термов.

После этого вектор приписывается к тому кластеру, к которому он наиболее близок. Далее векторы группируются и перенумеровываются соответственно принадлежности к кластерам. Потом для каждого из новых кластеров заново определяется центроид $c^i = (c_1^i,\ ...,\ c_N^i)$ – вектор, наиболее близкий ко всем векторам из данного кластера, координаты которого определяются, например, следующим образом:

$$c_k^i = \frac{1}{|C_i|}\sum_{x \in C_i} x_k.$$

После этого снова осуществляется процесс наполнения кластеров, затем вычисление новых центроидов и т.д., пока процесс формирования кластеров не стабилизируется (или если уменьшение суммы расстояния от каждого элемента до центра его кластера меньше некоторого заданного порогового значения).

Алгоритм *k*-means максимизирует функцию качества кластеризации Q:

$$Q(C_1,\ ...,\ C_k) = \sum_{i=1}^{k}\sum_{x \in C_i} Sim(x, c^j).$$

В отличие от метода LSI, *k*-means может использоваться для группирования динамических информационных потоков благодаря своей

вычислительной простоте – $O(kn)$, где n – количество объектов группирования (документов). Недостатком метода является то, что каждый документ может попасть всего лишь в один кластер.

3.3.2. Иерархическое группирование-объединение

Иерархическое группирование-объединение (Hierarchical Agglomerative Clustering, HAC) начинается с того, что каждому объекту в соответствие ставится отдельный кластер, а затем происходит объединение кластеров, которые наиболее близки друг к другу, в соответствии с выбранным критерием. Алгоритм завершается, когда все объекты объединяются в единый кластер. История объединений образует бинарное дерево иерархии кластеров.

Разновидности алгоритма HAC различаются выбором критериев близости (подобия) между кластерами. Например, близость между двумя кластерами может вычисляться как максимальная близость между объектами из этих кластеров.

Иерархическая кластеризация очень часто применяется при социологическом анализе, в биологии, экономике и т.д. Главным образом там, где заранее неизвестно количество кластеров.

Для иерархической кластеризации необходимо каким-либо образом определить расстояние между узлами рассматриваемого графа (сети). Т.е. нам необходимо получить количественную оценку близости узлов, аналогичную расстоянию в обычном евклидовом пространстве. Рассмотрим два наиболее используемых определения. Первое это Евклидово расстояние (Euclidean distance), определяется следующим образом:

$$x_{i,j} = \sqrt{\sum_{k \neq i,j}^{N} \left(A_{ik} - A_{jk} \right)^2},$$

здесь N – число узлов в сети. Евклидово расстояние точно равно нулю для полностью структурно –

эквивалентных узлов и увеличивается для узлов, которые не имеют общих соседей. Второе определение базируется на корреляции Пирсона между строками (столбцами) матрицы инцидентности:

$$x_{i,j} = \frac{\frac{1}{N} \cdot \sum_{k=1}^{N} (A_{ik} - \mu_i) \cdot (A_{ik} - \mu_i)}{\sigma_i \cdot \sigma_j},$$

где

$$\mu_i = \frac{1}{N} \cdot \sum_{j}^{N} A_{ij}, \quad \sigma_i^2 = \frac{1}{N} \cdot \sum_{j}^{N} (A_{ij} - \mu_i)^2.$$

При реализации алгоритма НАС могут использоваться различные меры близости, например, близость «центров масс», средняя близость между всеми парами объектов в объединенных кластерах и т.п. Мера близости между двумя кластерами C_i и C_j в последнем случае вычисляется по формуле:

$$Sim(C_i,\ C_j) = \frac{1}{|C_i \cup C_j|(|C_i \cup C_j| - 1)} \sum_{x,y \in C_i \cup C_j, x \neq y} Sim(x,\ y).$$

В этом выражении $|C_i \cup C_j|$ – количество объектов в множестве $C_i \cup C_j$, а x и y – объекты, принадлежащие $C_i \cup C_j$.

Сложность алгоритма НАС составляет $O(n^2 s)$, где n – количество объектов, а s – сложность вычисления близости между кластерами.

3.3.3. Латентно-семантический анализ

Метод кластерного анализа LSA/LSI (латентного семантического анализа/ индексирования) [базируется на сингулярном разложении матриц (SVD, Singular Value Decomposition). Пусть массиву документов $D = \{d_j \mid j = 1,\ ...,\ n\}$ ставится в соответствие матрица

A, строки которой соответствуют документам, а столбцы – весовым значениям термов (размер словаря термов – m). Сингулярным разложением матрицы A ранга r размерности $m \times n$ называется ее разложение вида $A = USV^T$, где U и V – ортогональные матрицы размерности $m \times r$ и $r \times n$, соответственно, а S – диагональная матрица, элементы которой $s_{ij} = 0$, если $i \neq j$, а диагональные элементы $s_{ii} \geq 0$. Диагональные элементы матрицы S называют сингулярными значениями матрицы A.

Ортогональные матрицы U и V обладают таким свойством:

$$UU^T = V^T V = I.$$

Доказано, что приведенное выше разбиение матрицы A обладает той особенностью, что если в матрице S оставить только k наибольших сингулярных значений (обозначим такую матрицу как S_k), а в матрицах U и V – только соответствующие этим значениям колонки (соответственно, матрицы U_k, V_k), то матрица $A_k = U_k \cdot S_k \cdot V_k^T$ будет наилучшей по Фробениусу аппроксимацией исходной матрицы A матрицей с рангом, не превышающим k. Напомним, что нормой матрицы X размерности $M \cdot N$ по Фробениусу является выражение:

$$\|X\|_F = \sqrt{\sum_{i=1}^{M} \sum_{j=1}^{N} x_{ij}^2}.$$

Таким образом, для матриц A и A_k доказано, что:

$$A_k = \underset{X : rank(X) = k}{\arg\min} \|A - X\|_F.$$

В соответствии с методом LSA в рассмотрение берутся не все, а лишь k наибольших сингулярных

значений матрицы A, и каждому такому значению ставится в соответствие один кластер.

A_k определяет k-мерное факторное пространство, на которое проецируются, как документы (с помощью матрицы V), так и термины (с помощью матрицы U). В полученном факторном пространстве документы и термины группируются в области, имеющие некоторый общий скрытый смысл. Т.е. получаемые области и представляют собой кластеры.

Выбор наилучшей размерности k для LSA – это проблема отдельных исследований. В идеале, k должно быть достаточно велико для отображения всей реально существующей структуры данных, но в то же время достаточно мало, чтобы не учитывать шума – случайных зависимостей.

В практике информационного поиска особое значение отводится матрицам U_k и V_k^T. Строки матрицы U_k рассматриваются как образы термов в $k-$мерном вещественном пространстве. Аналогично, столбцы матрицы V_k^T рассматриваются как образы документов в том же $k-$мерном пространстве. Иными словами, эти векторы задают искомое представление термов и документов в $k-$мерном пространстве скрытых факторов.

Существуют также методы инкрементного обновления всех значений, используемых в методе LSA. При пополнении новым документом d (например, новым результатом поиска по запросу) информационного массива, для которого уже проведено сингулярное разложение, можно не вычислять разложение заново. Достаточно аппроксимировать его, вычисляя образ нового документа на основе ранее вычисленных образов термов и весов факторов. Пусть $d-$вектор весов термов нового документа (новый

столбец матрицы A), тогда его образ можно вычислить по формуле: $d' = S_k^{-1} U_k^T d$.

Если q – вектор запроса пользователя размерности m, i – й элемент которого равен 1, если терм с номером i входит в запрос, и 0 – в противном случае, то образ запроса q в пространстве латентных факторов будет иметь вид: $q' = q^T U_k S_k^{-1}$.

В этом случае мера близости запроса q и документа d оценивается величиной скалярного произведения векторов q' и $V_k^T\{d\}$ (здесь $V_k^T\{d\}$ обозначает d – й столбец матрицы V_k^T).

При информационном поиске, в результате того, что отбрасываются наименее значимые сингулярные значения, формируется пространство ортогональных факторов, играющих роль обобщенных термов. В результате происходит «сближение» документов из близких по содержанию предметных областей, частично решаются проблемы синонимии и омонимии термов.

Метод LSA широко применяется при ранжировании выдачи информационно-поисковых систем, основанных на цитировании, в частности в алгоритме HITS, который будет рассмотрен ниже.

Наряду с тем, что метод LSA не нуждается в предварительной настройке на специфический набор документов, и он качественно выявляет латентные факторы, к его недостаткам можно отнести невысокую производительность (скорость вычисления SVD соответствует порядку $O(N^2 \cdot k)$, где $N = |D| + |T|$, D – множество документов, T – множество термов, k – размерность пространства факторов) и то, что он не предусматривает возможность пересечения кластеров, что противоречит практике. Кроме того, ввиду своей вычислительной трудоемкости метод LSA

применяется только для относительно небольших матриц.

3.4. Экстрагирование понятий

Экстрагирование понятий (Feature Extraction) является технологией, которая обеспечивает получение семантически важной информации из текстов, приведенной в структурированном виде. В качестве структуры могут применяться как относительно простые понятия (ключевые слова, персоны, организации, топонимы), так и более сложные, например, имя персоны, ее должность в конкретной организации и т.п.

Данная технология включает три основных метода:

а) Entity Extraction – извлечение слов или словосочетаний, важных для описания содержания текста. Это могут быть списки терминов по текущей области, персон, организаций, географических названий и т.д.;

б) Feature Association Extraction – исследование связей между извлекаемыми понятиями;

в) Event and Fact Extraction – извлечение сущности, распознавание фактов и событий.

Технология извлечения понятий, которая основана на применении специальных семантико-лингвистических методов, дает возможность получать приемлемую точность и полноту.

Следует отметить, что подходы к извлечению различных типов понятий из текстов существенно различаются как по контексту их представления, так и по структурным признакам. Так, для выявления принадлежности документа к тематической рубрике могут использоваться специальным образом составленные запросы информационно – поисковым языком, включающим логические и контекстные операторы, скобки и т.п.. Выявление географических названий (топонимов) предполагает использование

таблиц, в которых кроме шаблонов написания этих названий используются коды и названия стран, регионов и отдельных населенных пунктов.

Как один из примеров рассмотрим алгоритм обнаружения названий фирм в текстах документов. На вход системы поступает документ, который анализируется в процессе последовательного прочитывания. После этого текст документа сравнивается с шаблонами, соответствующими названиям известных фирм и если таковые присутствуют, то они помещаются в специальную таблицу «документ-фирма».

Также система извлечения понятий предполагает выявление первоначально неизвестных названий фирм на основании как шаблонов, так и результатов структурных исследований текста. При этом, в частности, используется таблица префиксов названий фирм, которая содержит такие элементы, как, например, «ООО», «ЗАО», «АО», «Компания» и другие.

Выявленные понятия могут быть основой для построения многопрофильных информационных портретов или интерактивных ситуативных карт (сетей, узлами которых являются понятия, а ребрами – информационные связи между ними), соответствующих запросам пользователей. Непосредственно по данным, представленным на ситуативной карте, которая отражает наиболее актуальные понятия (термины, тематические рубрики, топонимы, фамилии персон, названия компаний) возможно выявление взаимосвязей, то есть сами ситуативные карты могут служить исходными данными для построения сетей взаимосвязей понятий.

Таблицы взаимосвязей понятий строятся как статистические отчеты, отражающие близость (совместное появление в документах или близость по сопутствующим контекстам) отдельных понятий.

Это симметричные матрицы, элементы которых – коэффициенты взаимосвязей понятий. Эти

коэффициенты пропорциональны количеству документов входного информационного потока, которые удовлетворяют обоим понятиям, или количества значимых лексических единиц, употребляемых совместно с данными понятиями.

Таким образом, взаимосвязь понятий может быть оценена с помощью двух алгоритмов:

- совместного вхождения – путем расчета совместного вхождения понятий в одни и те же документы;
- контекстной близости – путем расчета корреляций наборов ключевых слов, входящих в документы, в которых упоминались данные понятия.

Рассмотрим формальное определение таблицы взаимосвязей понятий TVP', построенной с помощью первого алгоритма. Обозначим p_j $(j=1,...,M)$ – понятие, D – массив документов, $d^{(i)} \in D$ $(i=1, ..., N)$ – документ, P_j – подмножество D, соответствующее понятию p_j, $e_j^{(i)}$ – знак соответствия понятия документа:

$$e_j^{(i)} = \begin{cases} 1, & d^{(i)} \in P_j, \\ 0, & d^{(i)} \notin P_j. \end{cases}.$$

Можно определить уровень связи понятий p_j и p_k:

$$v_{j,k} = \sum_{i=1}^{N} e_j^{(i)} e_k^{(i)}.$$

Значение $v_{j,k}$ в совокупности образуют матрицу таблицы взаимосвязей понятий TVP'.

Для случая второго алгоритма, учитывающего контекстную близость, таблицу взаимосвязей понятий *TVP"* формально определим следующим образом. Обозначим $W_i = \{w_1^{(i)},\ ...,\ w_n^{(i)}\}$ – документа $d^{(i)}$.

Введем понятие профиля понятия p_j $(j = 1,...,M)$ как множество ключевых слов из документов, соответствующих этому понятию:

$$IP(p_j) = \bigcup_{d^{(i)} \in P_j} W_i.$$

Введем также понятие словаря системы $S = \{s_1,\ ...,\ s_K\}$ как множества ключевых слов, входящих в D, и вектора $\boldsymbol{t}^{(j)} = (t_1^{(j)},\ ...,\ t_K^{(j)})$ с элементами $t_i^{(j)}$, которые соответствуют профилям темы:

$$t_i^{(j)} = \begin{cases} 1, & s_i \in IP(p_j),\ i = 1,\ ...,\ K, \\ 0, & s_i \notin IP(p_j),\ i = 1,\ ...,\ K. \end{cases}$$

В этом случае уровень связи понятий p_j и p_k можно определить следующим образом:

$$\tilde{v}_{j,k} = \boldsymbol{t}_j \boldsymbol{t}_k = \sum_{i=1}^{K} t_i^{(j)} t_i^{(k)}.$$

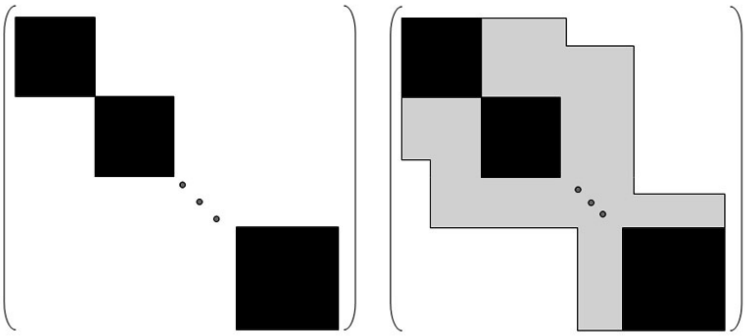
Следует отметить, что таблица взаимосвязей первого типа всегда отражает взаимосвязи понятий точнее, чем таблица взаимосвязей второго типа, однако таблица второго типа учитывает взаимосвязи более полно (рис. 30).

Действительно, из того факта, что $v_{j,k} > 0$, следует, что $\tilde{v}_{j,k} > 0$, поскольку первое условие определяет то, что существует хотя бы один такой документ (с индексом i), что $d^{(i)} \in P_j,\ d^{(i)} \in P_k$.

Отсюда следует, что сечение профайлов соответствующих понятий не является пустым:

$$IP\left(p_j\right) \cap IP\left(p_k\right) \neq \varnothing ,$$

и соответственно, $\vec{t}_j \vec{t}_k = \tilde{v}_{j,k} > 0.$



а) б)

Рис. 30 – Два варианта таблицы взаимосвязей понятий: нулевые элементы соответствуют белым областям, совпадающие – черным

Обратное утверждение в общем случае неверно.

Для переупорядочивания понятий в таблице взаимосвязей с целью выявления множеств наиболее взаимосвязанных с ними (путем выявления диагональных блоков) применяются методы кластерного анализа, в частности алгоритм *k*-means, который является одним из самых эффективных для группировки данных из динамических массивов.

3.5. Автоматическое построение аналитических отчетов

С самого начала компьютерной эры создавались программы автоматизированной обработки текстов, которые реализовали индексирование, аннотирование,

реферирование, фрагментирование и другие формы анализа и синтеза. Такие программы, с одной стороны, способствуют расширению информационного пространства (создавая новые документы – обобщения), а с другой – является почти единственным инструментом, который может обеспечить охват современных информационных ресурсов. Особенно большое значение приобрели задачи автоматического реферирования (Automatic Text Summarization) – составление кратких изложений материалов, аннотаций или дайджестов, т.е. извлечения наиболее важных сведений из одного или нескольких документов и генерацию на их основе лаконичных и информационно насыщенных отчетов. В настоящее время потребность в автоматическом реферировании текстов постоянно растет.

Вместе с тем нишу систем автоматического реферирования нельзя считать заполненной. Большинство процессов создания аннотаций еще неэффективны, появилась необходимость в масштабируемых методологиях и программах. Учитывая бурный рост технологий глубинного анализа текстов (Text Mining), ожидается большой прогресс также в области автоматического реферирования. Однако несмотря на то, что отдельные производители уже создали системы автоматического реферирования, объемы информации, которые порождаются в настоящее время, не позволяют оперативно получать аннотации необходимой полноты и релевантности.

Существует большое количество задач агрегирования текстов, которые достаточно четко подразделяются на два направления – квазиреферирования и краткого изложения содержания первичных документов. Квазиреферирование основано на экстрагировании фрагментов документов, извлечении наиболее информативных фраз и формировании из них квазирефератов.

Краткое изложение исходного материала основывается на извлечении из текстов с помощью

методов искусственного интеллекта и специальных информационных языков наиболее существенной информации и порождении новых текстов, содержательно обобщающих первичные документы. Применяя такой подход можно получать более сложные аннотации, которые в принципе могут содержать информацию, дополняющую исходный текст. Благодаря опоре на формальное представление семантики исходного документа, такие системы теоретически могут быть настроены на очень высокую степень сжатия, необходимую, например, для рассылки сообщений на мобильные устройства. Главное различие между средствами реферирования состоит в том, что они, по существу, формируют или набор выдержек, или краткое изложение.

Все существующие промышленные системы класса Text Mining включают средства автоматического реферирования, которые являются неотъемлемыми компонентами таких систем. Одна из базовых процедур систем этого класса – это автоматическое формирование дайджестов – автоматическое реферирование на основе большого количества документов. Для дайджеста отбираются документы, в которых наиболее явно отражены тенденции всего входного потока. Есть дайджесты, которые должны в наибольшей степени соответствовать информационным потребностям пользователя, по запросу которого формируется этот входной информационный поток.

Предполагается, что на основании реферата, который составляет по объему незначительную часть текста, пользователи смогут составить заключение относительно первоначального документа, затратив на это значительно меньше усилий. Как правило, при автоматическом реферировании объем реферата должен составлять от 5 до 30 % исходного текста. Подготовка документов, которые представляют собой аннотации из нескольких источников, то есть дайджестов, предполагает еще большую степень сжатия. При этом анализ качества реферирования – это отдельная очень важная задача, – зачастую не

предполагает однозначного решения. Как показывает практика, люди редко приходят к согласию относительно качества передачи основного содержания в реферате.

Несмотря на большую популярность методов искусственного интеллекта, в области автоматического реферирования в настоящее время можно констатировать то, что получение семантически наполненных результатов оказалось возможным и без привлечения баз знаний и правил. Вместе с тем разработчики средств автоматического реферирования все больше внимания уделяют гибридным системам, успешно сочетая статистические методы и методы искусственного интеллекта.

Большинство систем автоматического реферирования в настоящее время используют вариации статистических методов анализа, чаще игнорируя при этом лингвистическую взаимосвязь и семантику естественного языка. В таких системах автоматическое реферирование по сути является экстрагированием, т.е. квазиреферированием. Развитый синтаксический разбор и применение баз знаний или тезаурусов встречаются очень редко.

Вместе с тем в наиболее развитых системах реферирования учитывается зависимость предложений друг от друга, анафорические связи, обеспечивающие связность результирующих аннотаций, подбираются группы взаимосвязанных предложений, которые для большей связности несколько изменяются при аннотировании. Как правило, при этом предложения, характеризуются как «обрывки», например, начинающиеся со слов «При этом …», «Во-вторых …» чаще игнорируются подобными системами. Квазиреферирование сводится к экстрагированию (извлечению) из документов минимальных релевантных фрагментов. При этом оно имеет ту особенность по сравнению с кратким изложением, основанным на анализе поверхностно-синтетических отношений лексических единиц в тексте, выраженных в нем и не

требующих обращения к семантическим процессам, изученность которых недостаточна для описания свойств любого текста.

Второе направление формирования изложений на основе использования баз знаний, который является перспективным, в настоящее время представлен лишь экспериментальными исследованиями и до широкой реализации дело еще не дошло.

Квазиреферирование предполагает акцент на выделение характерных фрагментов методом сопоставления фразовых шаблонов, выделение крупнейшей блоков лексической и статистической релевантности. Автоматическое определение частот отдельных слов и сочетаний в исходном документе позволяет в итоге определять абзацы предложения, в которых тематика документа представлена точно. Создание итогового документа в этом случае – простое соединение выбранных фрагментов. Сформированный квазиреферат выглядит как связный текст, однако качество реферирования при этом во многом зависит от жанра текста. Гладкость и содержательность квазиреферата также зависит и от других особенностей исходного текста. Так, например, для больших текстов, монографий, интервью построение качественного реферата и фрагментов исходного документа без учета семантических закономерностей практически невозможно. Основу аналитического этапа квазиреферирования составляет процедура вычисления весовых коэффициентов для каждого блока текста в соответствии с такими характеристиками, как расположение этого блока в оригинале, частота появления в тексте, частота использования в ключевых предложениях, а также с другими показателями.

В рамках квазиреферирования выделяют три основных направления, которые применяются в современных системах совместно:

- статистические методы, основанные на оценке информативности различных элементов текста

частотностью, которая является основным критерием информативности слов, предложений или фраз;

- позиционные методы, опирающиеся на предположение того, что информативность элемента текста является зависимой от его позиции в документе;

- индикаторные методы, основанные на оценке элементов текста, исходя из наличия у них специальных слов и словосочетаний – маркеров важности (например, «в заключение», «было отмечено что …»), которые характеризуют их содержательную значимость. Существуют индикаторные методы, которые обеспечивают оценку фраз первичного документа на основе специальных словарей маркеров.

Следует отметить, что для русского языка, например, существуют словари маркеров, включающие более 1500 лексических единиц внетематичной лексики, а также формулы выбора, отражающие требования к вторичным документам, получаемых путем экстрагирования фраз на основе индикаторных методов. Эти элементы лексического аппарата обеспечивают достаточно точную идентификацию фрагментов исходного текста.

Большинство алгоритмов автоматического реферирования документов предполагают три основных шага: анализ исходного текста, определение значимых фрагментов (предложений или целых абзацев) и собственно формирование реферата.

Первый шаг начинается с выделения из исходного текста лексических единиц – слов или словосочетаний, их взвешивания по некоторым критериям и определения массива самых значимых. При этом сначала выполняется выделение всех лексических единиц из исходного текста и построение из них последовательного словарного массива. Затем каждой

лексической единице присваивается коэффициент, зависящий от ее расположения в исходном тексте. Затем выполняется их нормализация с помощью средств автоматического морфологического анализа (в настоящее время – это уже решенная проблема).

Морфологический анализ решает задачу приведения всех слов к каноническому виду. Цель морфологического анализа состоит из выделения основ слов, т.е. словоформ без флексий, а также при необходимости в подключении синонимических цепочек для отдельных слов. Для выполнения следующего семантического анализа каждой словоформе ставится в соответствие значение грамматических категорий (род, падеж, склонение и т.д.).

На этом этапе также выполняется удаление из словарного массива слов, не несущих явного смысловой нагрузки. Для этого применяются программные средства, основанные на использовании так называемого «стоп-словаря». Затем все лексические единицы массива сортируются и определяется их частотность в тексте. При этом каждой из лексических единиц приписывается весовой коэффициент, который определяется как результат учета нескольких составляющих: частотности, тематического словаря, обусловленного, например, тематикой запроса пользователя и «плюс-словаря», который включает наиболее важную лексику. Последний этап при формировании массива лексических единиц заключается в выборе некоторого ограниченного количества самых «весомых» сроков. Массив лексических единиц, кроме задачи автоматического реферирования, в дальнейшем может быть полезен при лингвистических исследований текста.

Определение веса фрагментов (предложений или абзацев) исходного текста выполняется по алгоритмам, разработанным еще в 60-70-е годы XX-го века, стали уже традиционными. Общий вес текстового блока на этом этапе определяется по формуле:

$$Weight = Location + KeyPhrase + StatTerm$$

Коэффициент *Location* определяется расположением блока в исходном тексте и зависит от того, где появляется данный фрагмент – в начале, в середине или в конце, а также используется ли он в ключевых разделах текста, например, в заключении.

Ключевые фразы (*KeyPhrase*) представляют собой конструкции – маркеры, которые формируют резюме, типа «в заключение», «в данной статье», «в соответствии с результатами анализа» и т.п. Весовой коэффициент ключевой фразы может зависеть также от оценочного термина, например, «отличный».

Статистический вес текстового блока (*StatTerm*) вычисляется как нормированная по длине этого блока сумма весовых значений терминов, входящих в него, – слов и словосочетаний. После выявления определенной, заданной коэффициентом необходимого сжатия, количества текстовых блоков с высокими весовыми коэффициентами, они объединяются для построения квазиреферата.

Конечно, преимущество методов квазиреферирования заключается в простоте их реализации. Выделение текстовых блоков, не учитывающих взаимоотношений между ними, часто приводит к формированию несложных рефератов. Некоторые предложения могут оказаться пропущены, в них могут встречаться слова или фразы, которые невозможно понять без предварительно пропущенного текста. Попытки решить эту проблему в основном сводятся к исключению таких предложений из рефератов. Реже делаются попытки разрешения ссылок с помощью методов лингвистического анализа. В ряде человеко-машинных подходов создаются специальные интерфейсы, с помощью которых можно определить наличие смыслового разрыва или «висящего» слова. Очевидно, что такой подход не годится для массовой обработки текстов.

На основе анализа нескольких документов строятся так называемые дайджесты. При составлении дайджестов методы автоматического реферирования

одного документа распространяются на массив из большого количества документов.

Вместе с тем дайджест можно рассматривать как аннотированный источник гиперссылок на документы, лежащие в его основе.

При формировании дайджестов методами квазиреферирования практически невозможно получить связный текст. Объединение рефератов каждого из документов неизбежно будет содержать избыточную несвязанную информацию.

Однако при условии составления автореферата, который содержит определенное количество анонсов входящих документов и разделен на подразделы в соответствии с этими документами, описанный выше метод оказывается вполне приемлемым.

Как и в случае квазиреферирования одного текстового документа на первом этапе формирования дайджеста происходит отбор наиболее значимых лексических единиц, входящих в массив выходных документов (входной информационный поток), строится словарь системы.

Выбор исходных документов из входного массива построения дайджеста осуществляется также с учетом их весовых значений. Вес каждого документа определяется с учетом нормированной по длине документа суммы весовых значений отдельных слов, входящих в этот документ. Этап выбора документов для дайджеста состоит из таких шагов, как определение веса каждого документа, сортировка входного потока документов по весовым значениям, определение смысловых дубликатов документов по статистическим критериям, исключение документов, непригодных для построения дайджестов (недопустимых типов документов, например, обзоров), а также смысловых дубликатов (определяют по специальным алгоритмам, например, частотным).

Последний этап выбора документов для

формирования дайджеста заключается в выборе заранее определенного количества самых весомых документов из отсортированного и отфильтрованного на предыдущих шагах массива. Статистический алгоритм обнаружения документов, дублирующихся из входного потока, может базироваться, например, на определении цепочек ключевых слов и частоты их появления для отдельных документов и в последующем сравнении между собой всех этих цепочек исходных документов.

Заключительный этап синтеза дайджеста заключается в выделении из выбранных документов самых значимых предложений и построении из них единого текста, разделенного на подразделы. Для этого к каждому из выбранных документов может применяться описанный выше алгоритм квазиреферирования.

Избранные документы присутствуют в дайджесте заранее заданным количеством весомых предложений. В комплексе формирования дайджестов на основе динамических текстовых потоков с Интернет, который реализован, например, в системе контент-мониторинга InfoStream, автоматически формируется гипертекстовое представление самого дайджеста, который можно рассматривать как самостоятельный документ, содержащий ссылки на документы-первоисточники в Интернет.

Приведенная выше процедура обеспечивает формирование дайджеста, который отражает основные тенденции, представленные в исходном информационном массиве. Вместе с тем имеет смысл формирования многоаспектного дайджеста, который отражает наряду с главной тенденцией несколько других аспектов, игнорируемых в дайджестах первого типа. В системе InfoStream реализовано построение именно многоаспектных дайджестов, которые строятся на базе технологических основ, применяемых при предыдущем подходе, путем применения алгоритма, который строится с учетом тематики запроса и многократной обработки информационного потока с изъятием на каждом шагу охваченных ранее

документов.

Подход, опирающийся на методы искусственного интеллекта, исходит из предположения, что если удается определить семантику текста, то аннотация будет более качественной. Базы знаний, применяемые при этом, должны постоянно поддерживаться в актуальном состоянии и сопровождаться экспертами. Для реализации этого метода необходимы многочисленные справочники, отражающие понятия, ориентированные на предметные области, необходимые для анализа и определения наиболее важной информации.

В отличие от частотно-лингвистических методов, обеспечивающих квазиреферирование, подход, основанный на базах знаний, опирается на автоматизированный качественный контент-анализ, который состоит, как правило, из трех основных стадий. Первая – отнесение исходной текстовой информации из заданного числа фрагментов (смысловых единиц) к определенным категориям. На второй стадии осуществляется поиск регулярных связей между смысловыми единицами, после чего наступает третья стадия – формирование выводов и обобщений. Строится структурная аннотация, которая представляет содержание текста в виде совокупности концептуально связанных единиц значения.

Семантические методы формирования рефератов-изложений предполагают два основных подхода: метод синтаксического разбора предложений и методы, опирающиеся на понимание естественного языка. В первом случае используются деревья разбора текста. Процедуры автоматического реферирования манипулируют непосредственно деревьями, выполняя перегруппировку и сокращение отраслей на основании структурных критериев. Такое упрощение обеспечивает построение автореферата – структурное «извлечение» исходного текста.

Второй подход основывается на системах искусственного интеллекта, в которых также на этапе анализа выполняется синтаксический разбор текста, но

синтаксические деревья не порождаются. В этом случае формируются семантические структуры, которые накапливаются в базе знаний. В частности, известны модели, позволяющие производить автоматическое реферирование текстов на основе психологических ассоциаций сходства и контраста. В базах знаний избыточная информация, которая не имеет прямого отношения к тексту, устраняется путем отсечения некоторых концептуальных подграфов. Затем информация подвергается агрегированию методом слияния графов или обобщения. Для выполнения этих преобразований выполняются манипуляции с логическими предположениями, выделяются определяющие шаблоны в текстовой базе знаний. В результате преобразования формируется концептуальная структура текста – аннотация, то есть концептуальные «выдержки».

Многоуровневое структурирование текста с использованием семантических методов позволяет подходить к решению задачи реферирования путем:

- изъятия малозначительных смысловых единиц. Преимуществом метода является гарантированное сохранение значимой информации, недостатком – низкая степень сжатия;
- сокращения смысловых единиц – замена их основной лексической единицей, выражающей основное содержание;
- гибридного способа, который заключается в уточнении реферата с помощью статистических методов, с использованием семантических классов, особенностей контекста и синонимических связей.

Хотя использование семантических методов при реферировании чаще приводит к потере некоторых второстепенных смысловых элементов, они не снижают качество – точность, компактность и связность реферата. По сравнению с традиционными подходами,

внедрение технологий Text Mining при анализе ресурсов Интернет уже сегодня обеспечивает, наряду с включением рабочих мест пользователей в динамическое информационное пространство, еще и получение за счет обратных связей оперативных количественных и качественных аналитических срезов, что раньше было практически невозможным.

Кроме того, растет объем мультимедийной информации в Интернет, который делает ее также очень важным объектом для обработки средствами реферирования. Технологии автоматического реферирования должны обрабатывать данные различного типа на этапах анализа и синтеза, реализуя интеграцию информации различного типа. Стоит заметить, что это направление находится лишь в самом начале своего развития, но уже достигнуты определенные успехи. Например, в системе Яндекс.Новости сегодня появилось группирование по сюжетам не только текстовых сообщений, но и фото-, аудио- и видеофайлов.

Хотя в настоящее время подходы, которые не предполагают использования методов искусственного интеллекта, еще доминируют, однако системы, основанные на экспертных системах, в ближайшее время смогут получить большее распространение в тех областях, для которых существуют разработанные лингвистические механизмы и базы знаний.

3.6. Компьютерная лексикография в аналитической деятельности

Компьютерная лексикография – часть компьютерной лингвистики. Традиционно лексикография рассматривалась как наука об упорядочении лексики – составления словарей, сам процесс создания словарей или как совокупность словарных произведений. Но это понимание лексикографии в настоящее время уже считается слишком узким. Еще в 1936 году Л.В. Щерба писал, что работа лексикографа «должна иметь научный характер

и отнюдь не сводиться к механическому сопоставлению каких-то готовых элементов» [Щерба, 1936]. Сегодня общепризнано, что лексикография – это самостоятельная научная дисциплина, имеющая свой предмет исследования, свои научные принципы, собственную теоретическую проблематику [Широков, 1998], [Широков, 2005].

В частности, в работе [Широков, 2011] сформулированы принципы, основанные на базе теорий семантических состояний и лексикографических систем, которые заставляют языковую субстанцию приобретать словарную форму.

Компьютерная лексикография может рассматриваться, с одной стороны, как одно из направлений компьютерной лингвистики, а с другой – как инструментально-ориентированная ветвь общей лексикографии, задачей которой является представление словарной информации в компьютерных системах и обеспечение ее функционирования в современном информационном пространстве.

Непосредственное отношение к становлению компьютерной лексикографии как отдельной научной ветви имеет так называемый «лексикографический эффект», который объясняет наличие системообразующих инвариантов лексикографических систем. Этот эффект можно охарактеризовать как феноменологический, поскольку он базируется на общих информационных свойствах систем и не привязан к их конкретной структуре. Наряду с этим, этот эффект был выявлен и сформулирован именно для лексикографических систем. В результате наблюдений и обобщения поведения различных систем был выявлен общий для всех известных процессов признак фундаментального характера [Широков, 2011], а именно то, что в процессе эволюции системы любой природы в ее структуре индуцируется некоторая подсистема относительно устойчивых дискретных сущностей, которые играют роль элементарных информационных единиц всей системы. При этом все остальные

феномены системы представляют собой определенным образом организованные комбинации этих информационных единиц.

Указанная выше подсистема обладает свойствами, родственными свойствам лексической системы естественного языка: она генерирует в своей структуре нечто вроде тезауруса и грамматики. При этом совокупности элементарных информационных единиц, как правило, обладают относительной стабильностью своих характеристик. Именно эти явления и составляют содержание лексикографического эффекта, который имеет значительный потенциал операциональности, позволяя выявлять в системах соответствующие комплексы элементарных информационных единиц.

В контексте данной работы лексикографический эффект выступает прежде всего как феноменологическая и методологическая основа выявления в качестве элементарных информационных единиц лексем, терминов, которые должны составлять основу построения таких подсистем естественного языка (или ее сегментов), таких как терминологические системы – словари, тезаурусы, онтологии.

Лексикографии отводится особое место при проведении аналитических исследований как инструменту нормирования лексики из предметной области, в частности, в задачах индексирования документов, реализации автоматического поиска и глубинного анализа данных, накопленных в массивах нормативно-правовой информации и т.д. Компьютерная лексикография при этом находит свою практическую реализацию прежде всего в терминологических системах.

Необходимость построения терминологических систем для применения в аналитической деятельности обусловливается необходимостью:

- корректного толкования терминов с целью предотвращения ошибок, различных смысловых коллизий;

- получения корректного толкования терминов;
- получения справочной информации по тем вопросам, на которые существуют ответы в нормативно-правовых актах;
- экспертного анализа терминологии относительно дублирования, противоречий, пробелов для дальнейшего их устранения.

Как инструмент, который часто используется в составе лингвистического обеспечения информационных систем, можно рассматривать тезаурусы [Добров 2009] – структурированные списки ключевых слов, предназначенных для однозначного представления концептуального содержания документов и запросов. Тезаурус упорядочивается так, чтобы установить прозрачные эквивалентны, гомографические, иерархические и ассоциативные связи между терминами. Тезаурус содержит:

- дескрипторы – слова и словосочетания, которые однозначно обозначают понятия из темы тезауруса;
- недескрипторы – слова и словосочетания, которые в естественном языке обозначают те же понятия, что и дескрипторы, или эквивалентные понятия;
- семантические связи (связи на основе смысловых значений) между дескрипторами и недескрипторами, а также между самими дескрипторами.

Проблема омонимичности в тезаурусе разрешается тем, что каждое ключевое слово представляется в контексте, который делает это слово однозначным. Для решения проблемы синонимичности один из синонимов избирается как дескриптор, а другим синонимам предоставляется статус недескрипторов. Только дескрипторы могут использоваться при индексировании и формулировке запросов, при этом недескрипторы помогают пользователям выбрать дескриптор. Если

установлено соответствие между идентичными понятиями в разных языках, пользователь многоязычного тезауруса может формулировать запросы родным языком и искать документы независимо от языка, на котором они были индексированы.

В качестве примера рассмотрим модель Функционального классификатора по вопросам государственной службы [Ланде, 1999], применяемый в качестве классификатора соответствующего документального массива нормативно-правовых актов. Функциональный Классификатор содержит отдельные термины (слова и словосочетания) и их определения, которые связаны с документами, в которых они определяются (содержатся). Термины связаны между собой парадигматическими связями. В качестве основы этого классификатора применялся тезаурус [ГОСТ, 1990], [ДСТУ, 2001], в который были включены такие типы лексических единиц, как:

- отдельные слова;
- именные словосочетания;
- лексически весомые компоненты сложных слов;
- аббревиатуры;
- сокращения слов и словосочетаний.

Словосочетания включались в словарь, если опорным словом в них было существительное и выполнялось одно из условий:

- значение словосочетания не вытекает из значений его составляющих;
- хотя бы одна из составляющих словосочетания не используется в составе других словосочетаний;
- словосочетание является устойчивыми;
- отдельные слова словосочетания имеют слишком широкое значение;

- разделение словосочетаний на отдельные составляющие ведет к потере важных для поиска парадигматических связей.

При построении функционального классификатора построение понятийных и словарных статей предусматривало, что лексическим единицам приписывались указатели, соответствующие стандарту ISO 2788 (табл. 1).

Табл. 1. Основные указатели по ISO 2788

Тип указателя	Значение	Аналог по ISO 2788
Ссылка от аскриптора к дескриптору	смотри	USE
Ссылка от дескриптора к синонимическому дескриптору или к аскриптору	синоним	UF (used for)
Ссылка от аскриптора к комбинации дескрипторов	Используй комбинацию	USE ... + ...
Ссылка от дескриптора к высшему дескриптору	выше	BT (broader term)
Ссылка от дескриптора к вышему родовому дескриптору	выше-род	BTG (broader term generic)
Ссылка от дескриптора к вышему родовому высшему дескриптору, который означает целое	выше-целое	BTP (broader term partitive)
Ссылка от дескриптора к вышему родовому низшему дескриптору	ниже	NT (narrower term)
Ссылка от дескриптора к низшему видовому дескриптору	ниже-вид	NTG (narrower term generic)
Ссылка от дескриптора к низшему видовому дескриптору, который означат часть	ниже-часть	NTP (narrower term partitive)
Ссылка от дескриптора к дескриптору, который связан ассоциативно	Ассоциация	RT (related term)

Указатель определяет связи между лексическими единицами или понятиями и является результатом выполнения таких операций, как:

- устранение неоднозначностей лексических единиц;
- установление отношений синонимии;
- выбор дескриптора, который отвечает за весь класс синонимии при индексировании;
- установление иерархических и ассоциативных отношений дескрипторов.

Множество компьютерных инструментальных средств лексикографии разделяется на: 1 – программы поддержки лексикографических работ и 2 – автоматические словари, тезаурусы, онтология, которая базируется на применении специальных алгоритмов и лексикографических баз данных [Amsler, 1982]

Существуют программные комплексы, которые сочетают свойства первой и второй групп, например, "Викисловарь" – лексикографическая среда Wiktionary, к описанию которого обратимся ниже.

Еще один такой пример – WordNet – электронный тезаурус/семантическая сеть, которая вместе с соответствующим программным обеспечением со свободной лицензией была разработана для английского языка в Принстонском университете США, а теперь нашла развитие для разных языков [Fellbaum, 2005]. Слова в ней организованы в синонимические группы (синсеты – синонимические ряды слов, которые выражают общее значение); группы связаны одна с одною отношениями антонимии, гиперонимии, гипонимии и др., т.е. это информационный ресурс, который отображает сложные отношения между лексическими единицами языка.

В WordNet словарными статьями являются синсеты – множества слов-синонимов, которые помечают некоторый концепт в заданном контексте.

Каждая словарная статья имеет толкование, которое не допускает неоднозначного понимания. Для синсета явно указывается часть речи и толкования. Каждое слово, которое входит в состав синсета, может дополнительно иметь ряд атрибутов, например, признака доминантности, ссылки типа «идиома», «близкое значение» и т.п. Для каждого слова может быть приведен пример его употребления в заданном контексте – определяется набор высказываний и фразеологизмов, также определяются толкования.

При формировании синсетов частотность употребления используется для упорядочения элементов синсета: выделяется «доминанта» – больше всего часто используемое нейтральное слово для выражения лексикализованного понятия (словосочетания) – и второстепенные элементы синсета, которые существенно уступают доминанте в частоте использования.

Статистико-комбинаторные характеристики контекстов применяются для выявления типичных для данного варианта слова схем сочетаемости, они заносятся в WordNet в виде перечней валентностей, которые задаются в формально-грамматическом, значностном и синтаксическом планах.

Европейские проекты EuroWordNet и BalkanNet обеспечивают работу из WordNet практически всеми основными европейскими языками. При этом связь разных языковых версий WordNet осуществляется через межъязыковый индекс (Inter Lingual Index – ILI), общий для всех версий. В настоящее время уже существует подход для построения многоязычного, в том числе и украиноязычного WordNet общего назначения (Распоряжение КМ Украины от 17 июля 2003 года № 415-р "Об утверждении плана мероприятий по созданию украинской лингвистической системы в сети Интернет: украинский вариант системы WordNet (UkrWordNet)")[Анисимов, 2005].

WordNet также может быть представлен как лексическая онтология – один из компонентов

Семантического веба – технологии W3C. Такое представление позволяет разным программным агентам интерпретировать данные из разных систем.

Семантические словари WordNet могут быть описанн средствами языка OWL (Ontology Web Language) и представлять один из ресурсов Семантического веба.

Проект EuroWordNet хранит основную базовую конструкцию WordNet, при этом обеспечивается расширение набора лексических отношений. К примеру, итальянский WordNet (ItalWordNet, IWN) сегодня состоит из 70 тыс. слов, организованных в 50 тыс. синсетов. Онтология в правовой сфере Jur-WN – это многослойный италоязычный лексикографический ресурс, который содержит большой набор семантических отношений (унаследованных от лингвистической конструкции общей базы данных IWN). Jur-WN представляет собой лингвистическую онтологию для правовой отрасли, которая обеспечивает возможности качественного поиска правовой информации (законодательства, судебных дел, политики) в многоязычных источниках. Jur-WN охватывает юридическую лексику (11 тыс. ключевых слов, 12 тыс. биграмм), что позволяет корректно обрабатывать информацию с учетом таких языковых явлений как полисемия и синонимия (содержит 2000 синсетов), обеспечивает взаимодействие с правовыми ресурсами пользователей, которые не являются юристами.

При анализе текстов, в частности, нормативно-правовых актов, необходимо в автоматизированном режиме особенным образом идентифицировать отличительные особенности единиц текста. Для этого важнейшим средством является конкорданс – лексикографическое произведение, которое является перечнем всех случаев употребления каждого слова в определенном тексте. Каждый случай словоупотребления дополняется информацией о контексте, о позиции лексической единицы, о ее

словесном окружении. Конкордансы могут использоваться для исследования соединений лексических единиц, нюансов значений, как источник для лексикографических примеров применений лексических единиц. С другой стороны, конкорданс является специализированной лингвистической прикладной программой, с помощью которой осуществляется автоматическая выборка заданных языковых единиц из электронных текстов. Функцией конкорданса является анализ одновременно нескольких текстов или корпусов электронных текстов, при этом конкорданс предоставляет пользователю информацию относительно контекста использования заданных языковых единиц. Конкорданс может предоставлять информацию о частотности употребления и сочетаемости той или другой языковой единицы, а также обращаться к тексту, в котором был найден пример.

Характеристики конкордансов зависят от таких параметров, как полнота описания, организация контекста, языковые или понятийные подходы и тому подобное.

Качественно различаются конкордансы типа KWIC (Keyword In Context) – ключевое слово в контексте и типа KWOC (Keyword Out of Context) – ключевое слово вне контекста [Дубичинский, 2008].

По-видимому, именно правовая наука, является «передовым рубежем» применения компьютерной лексикографии. Неопределенность, многозначность, даже противоречивость значений отдельных слов, терминов, понятий в нормативно правовых документах существенно усложняют условия существования, взаимодействия людей, общества, государства практически во всех отраслях жизни.

Особенность юридической лексики, в частности, заключается в использовании большого количества антонимов, потому что право регулирует интересы, которые отличаются своей противоположной направленностью. Кроме того, в правовой лексике

присутствует большое количество синонимов, которые в ряде случаев имеют разную смысловую нагрузку. Широко распространена также омонимия (одинаковые слова которые интерпретируются разным образом) и полисемия (когда одни и те же юридические термины имеют несколько разных значений).

К тому же добавляется потребность гармонизации нормативно-правовых документов нашего государства с соответствующими международными документами, что определяется современными мировыми интеграционными процессами. Именно для этого создан и адаптирован в Украине международный тезаурус EUROVOC – многоязычный политематический информационно-поисковый тезаурус, признанный как международный терминологический стандарт. Он реализован в соответствии со стандартами ISO 2788-1986 "Guidelines for the establishment and development of monolingual thesauri" ("Пособие по введению и разработке одноязычных тезаурусов") и ISO 5964-1985 "Guidelines for the establishment and development of multilingual thesauri" ("Пособие по введению и разработке многоязычных тезаурусов").

EUROVOC охватывает все основные темы, важные для деятельности европейских институций: политика, международные отношения, европейские содружества, законодательство, экономика, торговля, финансы, социальные вопросы, образование, коммуникации и тому подобное. Этот тезаурус реализован всеми официальными языками Европейского Союза.

EUROVOC имеет двухуровневую иерархию, верхний уровень которой складывают темы – им отвечают двухсимвольные коды. Нижний уровень организован как совокупность микротезаурусов (обозначенных четырьмя цифрами). Нумерация тем и микротезаурусов единственная для всех языков.

В программной реализации в среде Windows на экране системы EUROVOC одновременно представлены две панели, которые иллюстрируют выбранный уровень иерархии: список тем и микротезаурусов, или список

микротезаурусов и содержание выбранного микротезауруса, или микротезаурус и его отдельный дескриптор.

На уровне отдельных дескрипторов и недескрипторов структура EUROVOC зависит от семантических отношений, установленных между ними. Предусмотрены такие их типы:

– SN (Scope Note, примечание относительно возможных значений) – определение, которое уточняет значение дескриптора, или указание, как использовать дескриптор при индексировании документа и формулировке запросов;

– MT (Microthesaurus, микротезаурус) – ссылка на микротезаурус, к которой принадлежит дескриптор (недескриптор);

– UF (Used For, использованный для) и "USE" (использует) – связь эквивалентности между дескриптором и недескриптором (-ами), (UF) или между недескриптором и дескриптором, который заменяет этот недескриптор (USE). Фактически связь эквивалентности охватывает несколько типов связей:

- полной синонимичности или идентичного значения;
- близкой синонимичности или похожего значения;
- антонимии или противоположного значения;
- включение, когда дескриптор охватывает одно или больше понятий, которым предоставлен статус недескрипторов;
- иерархические связки между дескрипторами.

Существуют такие связки между дескрипторами:

BT (Broader Term, более широкий термин – между определенным и родовым (более обобщенным)

дескриптором – отмечается с числом, которое показывает количество шагов по иерархии между ними;

NT (Narrower Term, более узкий термин) – между родовым и видовым (более узким) дескриптором – отмечается с числом, которое показывает количество шагов по иерархии между ними;

RT (Related Term, взаимосвязанные термины) – ассоциативные связки между дескрипторами. Ассоциативная связь показывает, что существует другой релевантный дескриптор. Предусмотрены ассоциативные связи таких типов:

- причины и следствия;
- органа или инструмента;
- иерархии (поскольку, как отмечено выше, полииерархия не допускается, потерянные иерархические связи можно заменить ассоциативными);
- сопровождение;
- последовательности во времени или пространстве;
- вхождение в состав;
- характерной черты;
- объекта действия или процесса;
- расположение;
- подобия;
- антонимии.

Ассоциативные связи имеют такие существенные характеристики:

- они симметричны;
- они несовместимы с иерархическими связками – если два дескриптора связаны иерархией, между

ними нельзя установить ассоциативную связь и наоборот;

- между дескрипторами, которые имеют общий термин верхнего уровня, не может быть установлено ассоциативной связи.

Навигация по тезаурусу осуществляется с помощью ссылок. Дескриптор можно выбрать, набрав на клавиатуре первую букву его названия. Также в программном обеспечении EUROVOC реализован полнотекстовый поиск и поиск по ключевым словам.

Среди успешных попыток создания электронных лексикографических ресурсов правовой направленности еще можно назвать российский ресурс «Толковый словарь современной информационно-правовой лексики» (www.morepc.ru/informatisation/ dic.html), Словопедия – «Словарь сроков, употребляемых в действующем Законодательстве Украины» (http://slovopedia.org.ua/), ЛІГА:ЗАКОН: Терминологический словарь юриста (http://liga.biz-plan.com.ua/ resursyi/spravochniki-i-instrumentyi/ terminologicheskij-slovar-yurista.html), НАУ: Словарь законодательных терминов (http://zakon.nau.ua) и тому подобное.

3.7. Компьютерный анализ значимости терминов

На данное время актуальным является задание определения того, какие из важных структурных элементов текста оказываются информационно-значимыми, такими, которые определяют информационную структуру текста. Использование таких элементов как опорных слов позволяет формировать онтологию, тезаурусы, поисковые образы, в частности, при обработке нормативно-правовой информации. Такие элементы могут использоваться также для многих процедур, которые охватываются концепцией Text Mining, например, поиск подобных документов, выявление дубликатов, построение сниппетов, информационных портретов, идентификации таких компонент текста, как

коллокации, сверхфразовое единство [Yagunova, 2012] и тому подобное.

Ключевые слова для поиска в тексте, опорные слова для автоматического экстрагирования значимых фрагментов текстов или формирования автоматических рефератов, выбираются с учетом такого свойства слов, как "распознавательная" или дискриминантная сила [Ланде, 2012]. При анализе текстов по правовой тематике, в частности, при решении задания формирования электронной энциклопедии законодательства на основе анализа всего массива законодательных актов Украины, оценка дискриминантной силы отдельных слов имеет важнейшее значение.

Опорные слова могут выделяться путем применения некоторых шаблонов-маркеров (сигнальных слов), которые находятся в тексте, грамматического разбора текстов и построения синтаксических сетей слов, или на основе некоторых статистических признаков.

В качестве примеров сигнальных слов для нормативно-правовых актов можно привести такие: "под … понимается"; "к … принадлежат"; "к … относится"; "термин … означает"; "термин … обозначает" и т.п.

Статистические подходы базируются на том, что если слово относительно равномерно распределено по тексту документа, то оно вряд ли может использоваться для эффективного содержательного поиска или служить основой выбора какого-то значимого фрагмента, который может рассматриваться как некоторое сверхфразовое единство.

Большинство известных информационно-поисковых систем и систем классификации информации в той или иной степени основываются на использовании векторно-пространственной модели поиска (Vector-Space Model), и, соответственно, описанию данных [Salton, 1975].

В векторно-пространственной модели поиска для учета дискриминантной силы слов было введено понятие инверсной частоты появления слова в отдельных документах массива. Предложенный метод взвешивания слов имеет сегодня стандартное обозначение – TF IDF, где TF указывает на частоту появления слов в документе, а IDF – на величину, обратную к количеству документов в массиве, которые содержат данное слово (точнее, логарифм, монотонную функцию от этой величины).

В рамках этой модели документ описывается вектором в некотором евклидовом пространстве, в котором каждому терму, который используется в документе, ставится в соответствие его вес, который определяется на основе статистической информации о его появлении в отдельном документе или в документальном массиве.

Оценка неравномерности вхождения слов возможна и на основе чисто статистических, дисперсионных оценок. В работе [Ortuño, 2003] предложена такая оценка дискриминантной силы слова:

$$\sigma_i = \frac{\sqrt{\langle d^2 \rangle - \langle d \rangle^2}}{\langle d \rangle},$$

где: $\langle d \rangle$ – среднее значение последовательности $d_1, d_2, ..., d_n$, n – количество появлений слова t_i в информационном массиве.

Если обозначить координаты (номера) вхождения слова t_i в информационном массиве как $e_1, e_2, ..., e_n$, то $d_k = e_{k+1} - e_k$, $(e_0 = 0)$.

Для визуализации неравномерности вхождения слов в тексты в [Ortuño, 2003] была предложена технология спектрограмм, которые внешне напоминают штрих-коды товаров, но вместе с тем не позволяют рассматривать вхождения слов в разных

масштабах измерений, как это делается, например, в вейвлет-анализе.

Предложены и реализованы инструментальные средства, которые позволяют визуализировать плотность появления слова в тексте в зависимости от ширины окна наблюдения. Через веб-интерфейс соответствующей программы вводится текст и слово для анализа. В результирующей спектрограмме по горизонтали откладываются номера вхождения слов в тексте, а по вертикали – ширина окна наблюдения. Одному вхождению слова отвечает светло-серый цвет. Если в соответствующее окно наблюдения попадает несколько целевых слов, то оно закрашивается более темным оттенком. Эксперт – прикладной лингвист по внешнему вид сразу может определить меру равномерности вхождения в текст слова, которое анализируется [Ландэ, 2009].

В частности, коэффициенты неравномерности вхождения отдельных слов в подборке законодательных актов Украины, которые относятся к формированию и развитию информационного пространства государства (законы Украины "О доступе к публичной информации", "Об Основных принципах развития информационного общества в Украине на 2007-2015 годы", "О телекоммуникациях", "О защите персональных данных", "Об основах национальной безопасности Украины"), которые были рассчитаны автором, приведены в Табл. 2, а соответствующие спектрограммы – на рис. 31-15.

При расчете коэффициента w_i использовался искусственный прием, суть которого состоит в том, что исходный текст разбивался на фрагменты фиксированной длины по 500 слов, которые при расчетах TF IDF рассматриваются как отдельные фрагменты документов. Как видно, неравномерность вхождения отдельных слов, которая точно выражается в коэффициентах w_i и σ_i, может быть отображена в спектрограммах. Однако монотонность роста значений

w_i нарушается лишь в одном случае (слова «Безпека» и «Електронний»), что объясняется разными подходами, которые применяются для расчета w_i и σ_i частым появлением первого слова.

Табл. 2. Значение коэффициентов неравномерности для отдельных слов в подборке законодательных актов

Слово	Вхождения	w_i (TF IDF)	σ_i
Технології	46	53,62	1,99
Оприлюднення	33	53,98	2,21
Безпека	102	96,15	2,41
Електронний	50	85.24	3.22
Регулювання	220	129,92	3,62

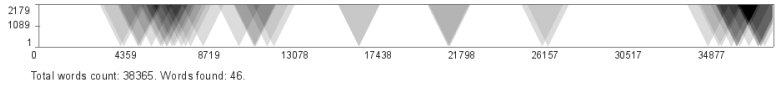


Рис. 31 – Спектрограмма вхождения слова «Технології»

Рис. 32 – Спектрограмма вхождения слова «Оприлюднення»

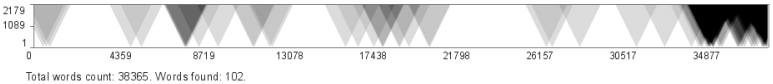


Рис. 33 – Спектрограмма вхождения слова «Безпека»

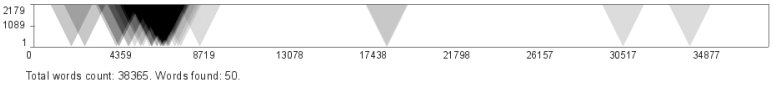


Рис. 34 – Спектрограмма вхождения слова «Електронний»

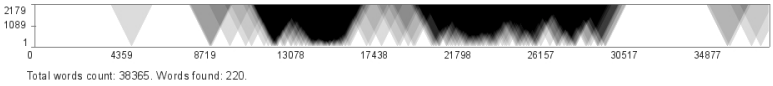


Рис. 35 – Спектрограмма вхождения слова «Регулювання»

Аналогичные расчеты были проведены для массива из 50 новостных публикаций веб в 2012 г. с тематикой, которая определяется запросом к системе мониторинга контента InfoStream (табл. 3, рис. 36-40):

(захист~персональн~даних) | кібербезпек | (інформац~безпек).

В этом случае монотонность роста значений по отношению к слову «Безпека» нарушается.

Следует обратить внимание, что дискриминантная сила отдельных слов на двух рассмотренных подборках существенно различается, что связано со стилем и содержанием соответствующих текстов.

Табл. 3. Значение коэффициентов неравномерности для отдельных публикаций

Слово	Входження	w_i (TF IDF)	σ_i
Регулювання	16	33,07	1,26
Оприлюднення	18	35,49	1,50
Безпека	56	71,59	1,75
Електронний	28	50.53	2.02
Технології	39	61,45	2,06

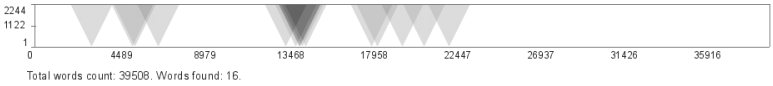


Рис. 36 – Спектрограмма вхождения слова «Регулювання»

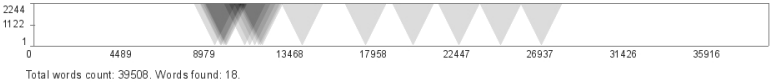


Рис.37– Спектрограмма вхождения слова «Оприлюдненння»

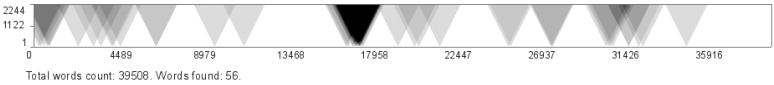


Рис. 38 – Спектрограмма вхождения слова «Безпека»

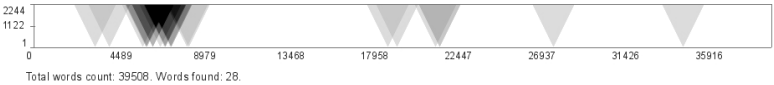


Рис. 39 – Спектрограмма вхождения слова «Електронний»

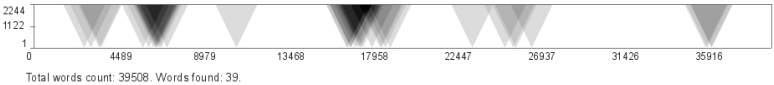


Рис. 40 – Спектрограмма вхождения слова «Технології»

Кроме традиционного подхода к оценке дискриминантной силы слов в текстах, предложенного Солтоном, дисперсионный анализ дает близкие по качеству результаты. Невзирая на то, что подход TF IDF за последнее время претерпел ряд трансформаций, дополняется вспомогательными параметрами, в частности, получил популярность метод BM25, который учитывает длину документов, дисперсионный анализ оказывается достаточно перспективным.

Рассмотренные примеры показали, что искусственный прием, который заключается в том, что исходный текст большого размера разбивался на фрагменты фиксированной длины, полностью оправдался, результаты во многом совпали с результатами, полученными другим методом.

Приведенные примеры показывают, что неравномерность слов в массивах новостных сообщений

и в официальных документах имеет близкую, во многом аналогичную природу, однако дискриминантная сила отдельных слов на двух рассмотренных подборках существенно различается, что связано со стилем и содержанием соответствующих текстов.

И, наконец, предложенный метод визуализации неравномерности вхождения слов, в сравнении с существующими, прибавил еще одно измерение – величину окна наблюдения, которое оказалось удобным при рассмотрении текстовых (в том числе документальных) массивов больших объемов. Техника спектрограмм позволяет экспертам без дополнительных усилий качественно оценивать значение отдельных слов при формировании так называемых сверхфразовых единств, экстрагировании фрагментов текстов для формирования справочных документов.

3.8. Сложные сети и задачи компьютерной лингвистики

3.8.1. Понятие сложных сетей

Информационные системы могут быть представлены как сетевые структуры, так называемые динамические сети [Newman, 2003], [Dorogovtsev, 2003]. Текущее состояние информационной системы может быть представлено в виде графа $< M, L >$, где M – это множество компонент (например, документов) информационной системы, а L – множество ребер, например, связей подобия, цитирования, ссылок и т.д. В настоящее время наряду с традиционными теориями графов, систем и сетей массового обслуживания активно развивается теория сложных сетей (от англ. – *Complex Networks*), в рамках которой предлагаются подходы к решению вычислительно сложных задач, характерных для современных сетей. Об актуальности теории сложных сетей свидетельствуют результаты современных работ по описанию реальных компьютерных, биологических и социальных сетей.

Такие сети имеют характеристики, не свойственные сетям с равновероятной связностью узлов,

и строятся на основе связных структур, степенных распределений и узлов-концентраторов.

Теория сложных сетей как область дискретной математики изучает характеристики сетей, учитывая не только их топологию, но и статистические феномены, распределение весов отдельных узлов и ребер, эффекты протекания, просачивания, проводимости в сетях тока, жидкости, информации и т.д. Оказывается, что свойства многих реальных сетей существенно отличаются от свойств классических случайных графов. Изучение таких параметров сложных сетей, как кластерность, посредничество или уязвимость напрямую относятся к теории живучести, так как именно от этих свойств зависит способность сетей сохранять свою работоспособность при деструктивном воздействии на их отдельные узлы или ребра (связи).

Важной характеристикой сети является функция распределения степеней узлов $P(k)$, которая определяется как вероятность того, что узел i имеет степень $k_i = k$. То есть распределение степеней $P(k)$ отражает долю вершин со степенью k.

Для ориентированных сетей существует распределение выходящей полустепени $P^{out}(k^{out})$, и полустепени входной $P^{in}(k^{in})$, а также распределение общей степени $P^{io}(k^{in}, k^{out})$. Последнее задает вероятность нахождения узла с входной полустепенью k^{in} и выходной полустепенью k^{out}.

Сети, характеризующиеся разными $P(k)$, демонстрируют весьма разное поведение. Распределение $P(k)$ в некоторых случаях может быть распределением Пуассона ($P(k) = e^{-m} m^k / k!$, где m – математическое ожидание), экспоненциальным ($P(k) = e^{-k/m}$) или степенным ($P(k) \sim 1/k^{\gamma}, \ k \neq 0, \ \gamma > 0$).

Важной особенностью многих реальных сетей

является распределение степеней узлов $P(k)$ по степенному закону.

Сети со степенным распределением степеней связности узлов называются безмасштабными (*scale-free*). Именно безмасштабными часто оказываются реально существующие сложные сети. При степенном распределении возможно существование узлов с очень высокой степенью, что практически не наблюдается в сетях с пуассоновским распределением.

Первым шагом при применении теории сложных сетей к анализу текста является представление этого текста в виде совокупности узлов и связей, построение сети языка (*Language Network*) [Головач, 2006].

Существуют различные способы интерпретации узлов и связей, что приводит, соответственно, к различным представлениям сети языка. Узлы могут быть соединены между собой, если соответствующие им слова стоят рядом в тексте [*Ferrer-i-Cancho, 2001*], [Dorogovtsev, 2001], принадлежат одному предложению [Caldeira, 2005], соединены синтаксически [*Ferrer-i-Cancho, 2004*], [*Ferrer-i-Cancho, 2005*] или семантически [Motter, 2002], [Sigman, 1999].

Сохранение синтаксических связей между словами приводит к изображению текста в виде направленной сети (directed network), где направление связи соответствует подчинению слова.

Поставим в соответствие каждому слову узел сети. Соединим каждые два узла связью, если соответствующие им слова стоят в предложении рядом. Такое представление называют *L*-пространством. В *L*-пространстве, равно как и в других приведенных ниже представлениях, при возникновении кратных связей принято сохранять лишь одну из них.

- *L*-пространство. Связываются соседние слова, которые принадлежат к одному предложению. Количество соседей для каждого слова (окно слова) определяется радиусом взаимодействия *R*, чаще всего рассматривается случай $R = 1$.

- *B*-пространство. Рассматриваются узлы двух видов, соответствующие предложениям и словам, которые им принадлежат.
- *P*-пространство. Все слова, которые принадлежат одному предложению, связываются между собой.
- *C*-пространство. Предложения связываются между собой, если в них употреблены одинаковые слова.

В случае *L*-пространства связи могут учитывать не только «ближайших соседей», но и группы из нескольких слов, которые находятся на определенном расстоянии друг от друга. Для этого вводится поняте «радиуса действия» R: при $R = 1$ связь существует лишь между ближайшими соседями, при $R = 2$ – между ближайшими и следующими близкими соседями и т. д. Переменная R может принимать значения от $R = 1$ до $R_{\max}$, где $R_{\max} + 1$ – общее количество слов в предложении.

Еще один способ представления текста в виде сети заключается в использовании двудольных (bipartite) графов. В таком представленнии (*B*-пространство) рассматриваются узлы двух видов. Один вид соответствует предложениям, второй – словам. Связь между различными узлами означает, что слово принадлежит предложению.

В *P*-пространстве все слова, принадлежащие одному предложению, считаются связанными между собой. В *C*-пространстве узлы соответствуют предложениям, а связь между узлами-предложениями устанавливается в том случае, если у них есть общие слова.

В случае рассмотрения *L*-пространства языка количество соседних слов, между которыми строятся связи, определяется параметром R: при $R = 1$ связаны между собой лишь ближайшие соседи, при $R = 2$ связи строятся между узлом-словом, его ближайшими и предшествующими близкими соседями и т. д. Рост

«радиуса взаимодействия» R приводит к росту количества связей, достигая насыщения при $R = R_{\max}$.

Для сети, построенной на базе Британского национального корпуса, оказалось, что данная сеть английского языка безмасштабна, а поведение степени $P(k)$ характеризуется двумя режимами степенного распределения со значением степенного показателя $\gamma = 1,5$ для $k < 2000$ и $\gamma = 2,7$ для $k > 2000$ соответственно.

Согласно определению, если средняя длина кратчайшего пути растет с размером (количеством узлов) сети медленнее любой функции степени, то сеть является «малым миром». Сети малого мира чрезвычайно компактны. Для упомянутой выше сети английского языка длина кратчайшего пути составляет всего $< l >= 2,63$. Поскольку рост R приводит лишь к добавлению новых связей, то $< l >$ уменьшается с ростом R.

Специфической формой корреляции в сетях является образование кластеров. Коэффициент кластерности C характеризует склонность сети к образованию соединенных троек узлов. Известно, что для полного графа $C = 1$, а для сети в форме дерева $C = 0$. Отношение среднего коэффициента кластерности исследуемых сетей к коэффициенту кластерности классического случайного графа свидетельствует о том, что сети языков являются хорошо коррелированными структурами. Такие корреляции растут с ростом «радиуса взаимодействия» R. Для Британского национального корпуса на основании анализа текстов, которые содержали $\approx 10^7$ слов, получено значение коэффициента кластерности $\langle C \rangle = 0,687$.

В случае рассмотрения $P-$пространства каждое слово-узел связано со всеми другими словами, которые принадлежат общему предложению. Таким образом, каждое предложение текста входит в сеть как полный

граф – клика (группа) взаимосвязанных узлов. Разные предложения-клики объединяются в сеть благодаря общим словам. В *L*-пространстве слова связываются в пределах окна, размеры которого характеризуются переменной R. Когда размер этого окна становится равным размеру предложения, то представление этого предложения в *L*- и в *P*-пространствах совпадают. Соответственно, когда размер окна становится равным размеру самого длинного предложения текста ($R = R_{\max}$), то представления всего текста в *L*- и в *P*-пространствах совпадают.

Получены результаты, которые убедительно свидетельствуют о том, что сеть языка яляется сильно коррелирующим безмасштабным малым миром (scale-free small world).

Имеется ряд научных трудов, в которых сделана попытка объяснить свойства сетей языка с помощью так называемого сценария подавляющего присоединения (preferential attachment), рассматривая их как результат процесса роста, когда новые узлы-слова с большей вероятностью присоединяются к узлам-хабам, имеющим наибольшее количество связей [Albert, 1999].

3.8.2. Сети горизонтальной видимости

В рамках концепции сложных сетей предложено несколько методов построения сетей на основе временних рядов, среди которых можно назвать несколько методов построения графов видимости [Nunez, 2012], в частности, так называемый граф горизонтальной видимости (Horizontal Visibility Graph – HVG)[Luque, 2009], [Gutin, 2011]. Эти подходы также позволяют строить сетевые структуры на основании текстов, в которых отдельным словам или словосочетаниям некоторым специальным образом поставлены в соответствие числовые весовые значения. Как функцию, которая ставит в соответствие слову число, можно рассматривать, например, порядковый номер уникального слова в тексте, длину слова,

общепринятую оценку TF IDF (в каноническом виде, равную произведению частоты слова в фрагменте текста – Term Frequency – на двоичный логарифм от величины, обратной количеству фрагментов текста, в которых это слово встретилось – Inverse Document Frequency) или ее варианты, а также другие весовые оценки.

При построении сетей слов в этой работе также используется дисперсионная оценка веса слов [Ortuño, 2003], формула для расчета которой была приведена выше.

Ряды из цифровых значений соответствующих слов можно превратить в графы горизонтальной видимости, в которых узлам соответствуют не только цифровые значения, но сами слова, которые имеют определенное содержательное значение. Сеть языка с использованием алгоритма горизонтальной видимости строится в три этапа [Lande, 2013]. На первом – на горизонтальной оси отмечается ряд узлов, каждый из которых отвечает словам в порядке появления в тексте, а по вертикальной оси откладываются весовые численные оценки (визуально – набор вертикальных линий, рис. 41).

На втором этапе строится традиционный граф горизонтальной видимости [Luque, 2009]. Для этого между узлами устанавливается связь, если они находятся в "прямой видимости", то есть если их можно соединить горизонтальной линией, которая не пересекает никакой вертикальной линии, размещенной между этими узлами.

Алгоритм построения графа горизонтальной видимости можно представить удобным для вычисления способом. Так например, на рис. 13 для узла-слова A_1^{n+2} (верхний индекс – номер слова в тексте, нижний – номер появления конкретного слова, в данном случае – слова A) смежными в сети считаются слова B_3^n и C_7^{n+5} и устанавливаются ребра-связи, такие что B_3^n –

ближайшее слева от A_1^{n+2} слово с весовой оценкой $\sigma_n = \sigma_B$, которая превышает весовую оценку слова A ($\sigma_{n+2} = \sigma_A$), а $C_7^m\,(m = n+5)$ – ближайшее справа от A_1^{n+2} слово, для которого $\sigma_{105} > \sigma_{102}$.

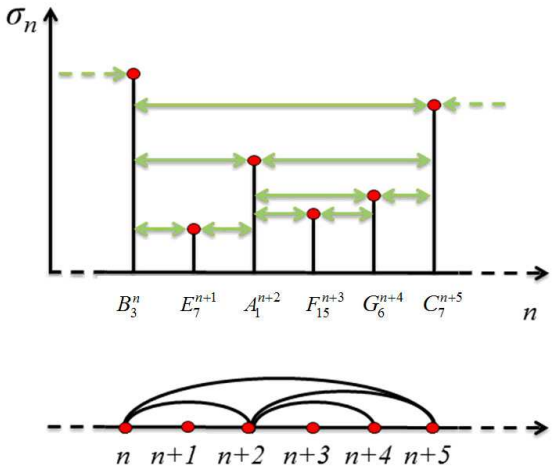


Рис. 41 – Пример построения графа горизонтальной видимости

На третьем, завершающем этапе, полученная на предыдущем этапе сеть компактифицируется. Все узлы с этим словом, например словом A, объединяются в один узел. Все связки таких узлов также объединяются. Важно отметить, что между любыми двумя узлами при этом остается не более чем одна связь – кратные связки изымаются. В частности это значит, что мера (число связей) узла не превышает сумму степеней $\sum_k A_k^n$.

В итоге получается новая сеть слов – компактифицированный граф горизонтальной видимости (КГГВ) – рис. 42.

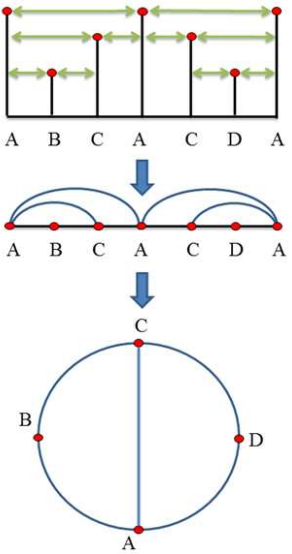


Рис. 42 – Этапы построения компактифицированного графа горизонтальной видимости

Для КГГВ-сетей слов было определено распределение степеней узлов, которое оказалось близким к степени ($P(k) = Ck^{\alpha}$), т.е. эти сети являются безмасштабными. Были проведены расчеты параметров сетей для многих текстовых документов. В результате оказалось, что для всех из них коэффициент α изменялся в диапазоне от 0,95 до 1,05.

В состав узлов с наибольшими степенями в КГГВ-сети, вместе с личными местоимениями и другими служебными словами (части, предлоги, союзы и все такое), попали слова, которые определяют информационную структуру текста [Giora, 1983], [Ягунова, 2010].

Для сравнения было дополнительно исследовано поведение простых сетей языка, когда на первом этапе построения сети связываются соседние слова, которые входят в текст (L-пространство, R = 1), а на втором происходит компактификация сети. Очевидно, вес

узлов в этой сети отвечает частоте появления слов, а их распределение – закону Ципфа [Zipf, 1949]. При этом наибольшие степени имеют узлы, которые отвечают словам с наибольшей частотой, – союзам, предлогам, местоимениям и тому подобное, что имеют большое значение для связности текста, но являются малоинтересными с точки зрения исследования информационной структуры.

Если обозначить Ψ – множество из N различных слов (например, $N = 100$), которые соответствуют самым весомым узлам приведенной простой сети слов, а Λ – множество слов, которые соответствуют самым весомым узлам КГГВ, то множество $\Omega = \Lambda \setminus \Psi$ отвечает информационно значащим словам, которые имеют, кроме того, важное значение и для связности текста.

В качестве примера сопоставлялись 100 наиболее весомых узлов для трех данных типов сетей слов за текстами Законов Украины "О телекоммуникациях" и "О защите персональных данных".

В КГГВ-сети по тексту Закона Украины "О телекоммуникациях" с учетом значений TF IDF в состав множества Ω попали такие слова, как «Державне», «Регулювання», «Ринку», «Інтернет», «Провайдер», «Трафік».

В КГГВ-сети для этого же текста по весовым значениям слов, соответствующих дисперсионным оценкам, дополнительно в множество Ω попали такие слова, как «Суб'єкт», «Ресурс», «Переоформлення», «Рішення», «Споживачів» и др.

При анализе текста Закона Украины "О защите персональных данных" в множество Ω (для КГГВ-сети с учетом весовых значений слов по алгоритму TF IDF) попали такие слова, как «Інформація», «Відстрочення», «Орган», «Баз», «Виключено».

В КГГВ-сети для текста настоящего законодательного акта по весовым значениям слов, которые отвечают дисперсионным оценкам, в

множество Ω попали дополнительно такие слова, как «Використання», «Прав», «Уповноважений», «Особа».

3.8.3.Онтологии

Термин «онтология» употребляется в нескольких областях знаний и имеет два разных значения:

- философская дисциплина, которая изучает наиболее общие характеристики бытия и сущностей;
- в инженерии знаний: артефакт, структура, которая описывает значение элементов некоторой системы.

В философии онтологией называют теорию о природе бытия и видах сущностей. В инженерии знаний онтологический уровень формализирует накопленные знания, определяя и сочетая терминологию разных предметных сфер. Таким образом, четкой взаимной обусловленности между значениями термина «онтология» в философии и в инженерии знаний не прослеживается. Связь между ними носит скорее ассоциативный характер.

Невзирая на существование большого количества наработок в отрасли представления знаний, не существует единственного четкого определения онтологии. Под онтологией в рамках данной работы будем понимать систему понятий предметной области, которая представляется как набор сущностей, которые объединены разными отношениями.

Онтологии получили широкое распространение в задачах представления знаний, семантической интеграции информационных ресурсов, информационного поиска и т.п. В науке об искусственном интеллекте онтология – это «спецификация концептуализации предметной области», или упрощенно, документ или файл, который формально задает связки между понятиями. Это своего рода словарь понятий предметной области и совокупность явно определенных предположений относительно

содержания этих понятий. Чаще всего онтология представляется как иерархия понятий, связанных отношениями определенных видов. Развитая онтология формализируется средствами логики и допускает возможности формирования логических утверждений.

Онтологии используются для формальной спецификации понятий и отношений, которые характеризуют определенную область знаний. Преимуществом онтологии как способа представления знаний является их формальная структура, которая упрощает компьютерную обработку.

Термину "онтология" удовлетворяет широкий спектр структур, которые представляют знания о той или другой предметной области. Так к онтологии можно отнести ряд структур, которые отличаются разным уровнем формализации:

- глоссарий;
- простая таксономия;
- тезаурус (таксономия с терминами);
- понятийная структура с произвольным набором отношений;
- полностью аксиоматизируемая теория.

В общем виде структура онтологии представляет собой набор элементов четырех категорий:

- понятие;
- отношение;
- аксиомы;
- отдельные экземпляры.

Исследователи выделяют прикладную онтологию, онтологию области знания, общую (род) онтологию и репрезентационную онтологию (идет речь об онтологии метауровня, которая включает в себя репрезентационные первоэлементы).

Онтологии могут быть также разделены на одноязычные и многоязычные. Уже существует ряд онтологий, ориентированных на представление знаний на нескольких языках, например, EuroWordNet, MikroKosmos и некоторые другие.

Также выделяется особенный тип онтологии – лексическая онтология (или лингвистическая). Отличительным свойством такой онтологии является "фиксация в одном ресурсе понятий (слов) вместе с их языковыми свойствами". Такая онтология тесно взаимоувязана с семантикой грамматических элементов (слов, именных групп и др.) Основными источниками понятий в онтологии данного типа являются значения языковых единиц. Их также отличает своеобразный набор отношений, обычно свойственный для языковых элементов: синонимия, гипонимия, меронимия, а также ряд других. К лингвистическим онтологиям авторы [Соловьев, 2006] относят WordNet, MikroKosmos, Sensus, РуТез и другие. Круг задач, решаемых такой онтологией, тесно взаимоувязан с обработкой естественного языка.

Онтологию, в частности, можно эффективно использовать для повышения точности информационного поиска – поисковая система будет выдавать только такие документы, где нужное понятие воспроизводится точно, а не те, в текстах которых встретилось заданное ключевое слово.

В Стенфордском университете (США) разработана программная платформа – редактор онтологии Protégé (http://protege.stanford.edu), а также организовано содружество энтузиастов, которое насчитывает несколько тысяч участников, которые пополняют базу онтологии для самих разных предметных областей. Заслуживает также внимания проект Estrella (www.estrellaproject.org), в рамках которого разработана онтология LKIF (Legal Knowledge Interchange Format) – язык для представления юридических знаний и обмена между правовыми информационными базами.

В настоящее время для разработки систем, основанных на знаниях, является актуальной задача

объединения разных репрезентативных подходов с целью обеспечения наиболее полного представления знаний в правовой сфере. В рамках разработки базовой правовой онтологии LKIF-Rus на базе международной онтологии LKIF-Core (LKIF – Legal Knowledge Interchange Format) была создана онтология гражданского права LKIF-CivilRus.

Современные средства онтологического моделирования позволяют частично внедрить промышленный подход в процесс разработки онтологии. Для этого, например, можно использовать SWRL-правила (SWRL – Semantic Web Rule Language), поддержка которых включена в среду разработки Protégé.

В рамках онтологии LKIF-CivilRus разработана группа SWRL-правил, которые регулируют институт действительности соглашений в гражданском праве – в зависимости от дееспособности и воли субъектов, а также соблюдения установленной законом формы соглашения.

Онтологическое исследование основ криминального права и разработка обобщенной онтологии этой предметной области LKIF-CrimRus происходило путем расширения русскоязычной онтологии верхнего уровня для системы русского права LKIF-Rus, в свою очередь основанной на базовой юридической онтологии LKIF-Core. Разработанная онтология основана на правовых нормах, которые содержатся в части первой Уголовного кодекса Российской Федерации. Онтология разработана с помощью онтологического редактора Protégé 3.4.4. Языком описания онтологии в Protégé является OWL. Работа из SWRL-правилами реализована в плагине SWRLTab для Protégé с использованием аппарата логического вывода Jess Rule Engine. Для визуализации онтологии был использован плагин Ontoviz на основе генератора диаграмм Graphviz. Для выполнения запросов к онтологии используется язык запросов SPARQL.

Формализация правовых норм является важным средством достижения совпадения или корреляции словаря конкретного субъекта квалификации преступлений с унифицированным словарем терминов криминального законодательства. В частности, онтология, которая охватывает ключевые понятия и категории всех норм криминального законодательства [Воронина, 2010], обеспечивает понятийную совместимость и единство информационно-поискового языка разной онтологии в области квалификации преступлений, индексирования в процессе квалификации данных и поиск необходимой информации для оценки действий лиц, которые совершили общественно опасные деяния.

На сегодняшний день существует вариант базовой правовой онтологии для системы русского права, которая имеет 8 уровней иерархии и включает 127 классов и 108 отношений. При адаптации онтологии LKIF-Core для русской правовой системы были заимствованы основные абстрактные концепции: часть-целое, пространственно-временные отношения, классификация материальных объектов.

Из правовых инструментов использована система правовой квалификации с помощью таких сущностей, как "Суждение", "Отношение к Суждению", "Квалификация", "Квалифицированный". Наиболее явные отличия в правовых системах были выявлены при детализации таких концептов, как "Источник" и "Субъект". Наследники этих классов определены исключительно российской теорией государства и права.

Существуют некоторые особенности использования онтологии для представления юридических знаний:

1. По мере развития любой правовой системы в нормативные акты вводятся новые или удаляются предыдущие причинно-следственные связки, что может привести к необходимости переопределения терминов, изменения их положения в таксономии. Таким образом

онтологию необходимо постоянно изменять. В связи с этим в онтологическом моделировании есть целое направление – управление версиями (versioning).

2. Разработчик онтологии не может гарантировать, что определение полностью отобразит смысл юридического понятия (по крайней мере, если это определение не из нормативного акта).

3. Не всегда возможным является выражение причинно-следственных сущностей правовых явлений.

4. Противоречие между требованием однозначного определения терминов в рамках правовой информации и практикой приводит к несогласованности онтологии, что недопустимо.

Среди проблем, которые возникают в процессе разработки международной правовой онтологии, наиболее существенными можно назвать следующие:

1) отличия в правовых системах и, как следствие, в правовом понятийном аппарате;

2) многозначность некоторых терминов, синонимия;

3) проблемы при обозначении отношений и особенно обратных отношений.

3.8.4. Сеть естественных иерархий терминов

Для решения задачи построения терминологической онтологии предметной области требуется проведение комплексных исследований, определенным этапом которых является построение так называемых словарных номенклатур, предметных словарей, тезаурусов. Эффективный автоматический отбор отдельных терминов для таких конструкций – не решенная окончательно задача, проблема установления связей, автоматического построения сетей из таких терминов до сих пор остается открытой.

В качестве терминологической основы для формирования онтологии предлагается использовать

сеть естественной иерархии терминов, которая базируется на информационно-значимых элементах текста [Yagunova, 2012], методология выявления которых приведена в [Lande, 2013]. Опорные термины, как правило, выбираются с учетом такого свойства, как дискриминантная сила. Однако одного этого свойства часто недостаточно для качественного отражения содержания предметной области. Иногда слова с низкой дискриминантной силой, например, наиболее частотные слова из выбранной предметной области (например, слова «Android», «IOS», «Приложение» в корпусе текстов по тематике современных гаджетов) оказываются важнейшими для рассматриваемой задачи.

Как подход к решению актуальной задачи построения терминологической онтологии, в данной работе рассматриваются принципы и методика формирования сети естественных иерархий терминов (СЕИТ), базирующейся на контенте научно-популярных статей выбранной направленности [Ландэ, 2014]. "Естественность" иерархий терминов в этом случае понимается как отказ при формировании сети от методов смыслового анализа текстов, ограничиваясь фактически статистическим анализом. Связи в такой сети определяются естественным взаимным положением слов и словосочетаний из текстов. Такая сеть, создаваемая полностью автоматически, может рассматриваться в качестве основы для дальнейшего автоматизированного формирования терминологической онтологии с участием экспертов.

Методика формирования сети естественных иерархий терминов, рассматриваемая в данной работе, предусматривает реализацию последовательности этапов, которые рассмотрим подробно:

1. На первом этапе формируется исходный текстовый корпус. Как пример такого корпуса рассматриваются полные тексты научно-популярных статей, опубликованных на веб-сайте «Компьютерра онлайн» (http://www.computerra.ru), посвященных

проблематике мобильных устройств, представленных на русском языке. В состав текстового корпуса было включено около 230 статей общим объемом свыше 800 тыс. символов. Предварительная обработка такого текстового корпуса предусматривала выделение фрагментов текстов (отдельных статей, абзацев, предложений, слов), исключение нетекстовых символов, отсечение флективных окончаний.

2. На втором этапе каждому отдельному термину из текста (слову, биграмме или триграмме) ставится в соответствие оценка их "дискриминантная сила", а именно TFIDF, которая в каноническом виде равна произведению частоты соответствующего термина (Term Frequency) в фрагменте текста на двоичный логарифм от величины, обратной к количеству фрагментов текста, в которых этот термин встретился (Inverse Document Frequency). Для последовательностей терминов и их весовых значений по TFIDF строятся компактифицированные графы горизонтальной видимости (КГГВ) и выполняется переопределение весовых значений слов уже по этому алгоритму. Данная процедура позволяет учитывать в дальнейшем кроме терминов с большой дискриминантной силой также высокочастотные термины, которые имеют большое значение для общей тематики. В соответствии с [Lande, 2013], сеть слов с использованием алгоритма горизонтальной видимости строится также в три этапа. На первом этапе на горизонтальной оси отмечается ряд узлов, каждый из которых соответствует словам в порядке появления в тексте, а по вертикальной оси откладываются весовые численные оценки (TFIDF). На втором этапе строится традиционный граф горизонтальной видимости [Luque, 2009]. Для этого между узлами существует связь, если они находятся в «прямой видимости», т.е. если их можно соединить горизонтальной линией, не пересекающей никакую другую вертикальную линию. На третьем, заключительном этапе, сеть компактифицируется. Все узлы с одним и тем же словом объединяются в один узел, связи таких узлов также объединяются. В качестве

весовых оценок отдельных слов в дальнейшем используются степени соответствующих им узлов в КГГВ. После этого все термины текста сортируются по убыванию рассчитанных весовых значений соответствующих узлов КГГВ. Дальнейшему анализу не подлежат термины из так называемого стоп-словаря, являющиеся важными для связности текста, но не несущие информационной нагрузки. Это, как правило, фиксированный набор служебных слов. Используемый в рамках данной работы стоп-словарь был построен на основе различных стоп-словарей, представленных в доступном виде на веб-ресурсах:

• *http://code.google.com/p/stop-words/source/browse/trunk/stop-words/stop-words/stop-words-russian.txt?spec=svn3&r=3;*

• *https://github.com/punbb/langs/blob/master/Russian/stopwords.txt;*

• *http://www.ranks.nl/stopwords/russian.html;*

• *http://trac.mysvn.ru/punbb/punbb/browser/trunk/Russian/stopwords.txt.*

Экспертным методом определяется необходимый размер СЕИТ (число N), после чего выбирается соответствующее количество единичных слов, биграмм и триграмм (всего $N + N + N$ элементов) с наибольшими весовыми значениями по КГГВ.

3. На третьем этапе из отобранных терминов строятся сети естественных иерархий терминов, в которых как узлы рассматриваются сами термины, а связи соответствуют вхождением одних терминов в другие. На рис. 43 проиллюстрирован принцип построения связей СЕИТ. Различные геометрические фигуры на этой иллюстрации соответствуют различным словам. Первой строке соответствует выбранное множество единичных слов, второй – множество биграмм, а третьей – множество триграмм. Если единичное слово входит в биграмму или триграмму, или биграмма входит в триграмму, образуется связь,

которая обозначается стрелкой. Множество узлов, которым соответствуют термины, и связи образуют трехуровневую сеть естественной иерархии терминов.

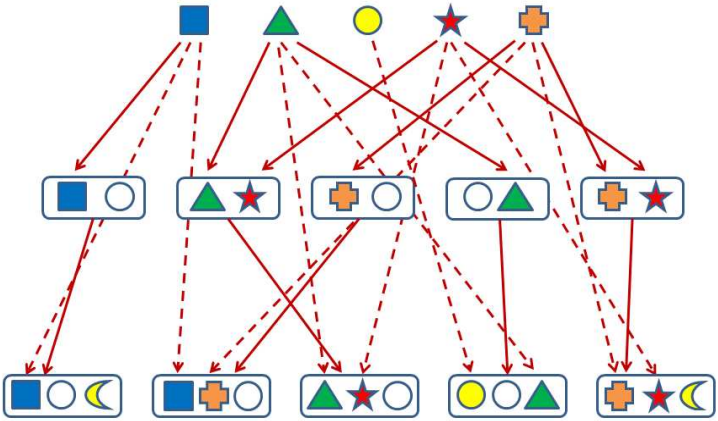


Рис. 43 – Трехуровневая сеть естественной иерархии терминов

После формирования СЕИТ (построения матрицы инцидентности) осуществляется ее отображение программными средствами анализа и визуализации графов. Для загрузки сетей естественных иерархий терминов в базы данных формируется матрица инцидентности общепринятого формата csv размерностью $(N + N + N) \times (N + N + N)$ элементов.

На рис. 44 представлена небольшая сеть естественной иерархии терминов размером 30+30+30, которая визуализирована средствами системы Gephi (https://gephi.org/).

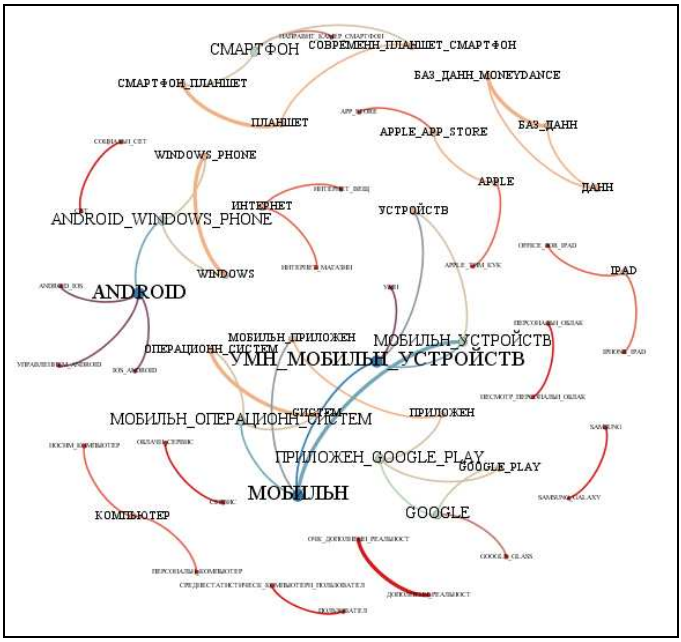


Рис. 44 – Визуализация связного фрагмента СЕИТ размером 30+30+30

На рис. 45 приведены отдельные фрагменты более крупной сети естественной иерархии терминов размером 200+200+200.

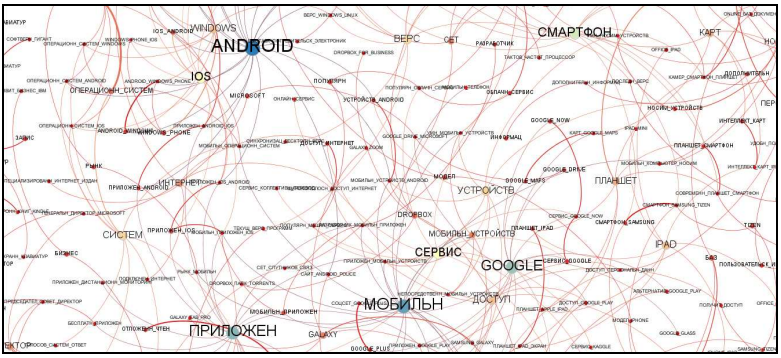


Рис. 45 – Фрагмент СЕИТ размером 200+200+200

Ранжирование узлов СЕИТ

Ранжирование узлов в СЕИТ возможно также по свойствам, определяемым сетевой структурой, связями. Например, для определения авторитетности узла как слова – источника порождения словосочетаний или как составного термина, состоящего из отдельных важных слов, можно анализировать СЕИТ, выбирая при этом наиболее важных «авторов» или «посредников». Для решения этой задачи предлагается использовать известный алгоритм ранжирования веб-страниц, основанных на связях, HITS (hyperlink induced topic search), предложенный Дж. Клейнбергом [Kleinberg, 1998].

Алгоритм ранжирования HITS обеспечивает выбор из информационного массива лучших «авторов» (узлов, на которые введут ссылки) и «посредников» (узлов, от которых идут ссылки цитирования). Понятно, что термин является хорошим посредником, если от него идут связи на важные словосочетания, и наоборот, термин (словосочетание) является хорошим автором, если на него ведут связи от важных авторов. В соответствии с алгоритмом HITS в нашем случае для каждого узла сети v_j рекурсивно вычисляется его значимость как автора $a(v_j)$ и посредника $h(v_j)$ по формулам:

$$a(v_j) = \sum_i h(v_i);\ h(v_j) = \sum_i a(v_i).$$

В данных формулах суммирование производится по всем узлам, которые ссылаются (или на которые ссылаются – во второй формуле) на данный узел.

Наиболее интересными с семантической точки зрения в рассматриваемой СЕИТ оказались узлы с наибольшим значением авторства и посредничества. В Табл. 4. приведены термины, соответствующие таким узлам.

Табл. 1. Термины, соответствующие узлам с наибольшим авторством и посредничеством

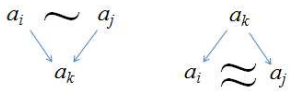
№	Термины с наибольшим значением посредничества	Термины с наибольшим значением авторства
1	МОБИЛЬНЫЙ	ПРИЛОЖЕНИЕ МОБИЛЬНОГО УСТРОЙСТВА
2	ANDROID	МОБИЛЬНОЕ ПРИЛОЖЕНИЕ IOS
3	IOS	МОБИЛЬНАЯ ОПЕРАЦИОННАЯ СИСТЕМА
4	СИСТЕМА	ОПЕРАЦИОННАЯ СИСТЕМА ANDROID
5	УСТРОЙСТВО	ПОЛЬЗОВАТЕЛЬ МОБИЛЬНОГО УСТРОЙСТВА
6	ОПЕРАЦИОННАЯ	МОБИЛЬНОЕ ПРИЛОЖЕНИЕ
7	ОПЕРАЦИОННАЯ СИСТЕМА	УМНОЕ МОБИЛЬНОЕ УСТРОЙСТВО
8	МОБИЛЬНОЕ УСТРОЙСТВО	ОПЕРАЦИОННАЯ СИСТЕМА IOS
9	WINDOWS	ПРИЛОЖЕНИЕ ANDROID
10	ПОЛЬЗОВАТЕЛЬ	ОПЕРАЦИОННАЯ СИСТЕМА WINDOWS
11	УМНЫЙ	МОБИЛЬНОЕ УСТРОЙСТВО
12	ПРИЛОЖЕНИЕ ANDROID	МОБИЛЬНЫЙ ПОЛЬЗОВАТЕЛЬ
13	МОБИЛЬНОЕ ПРИЛОЖЕНИЕ	ПРИЛОЖЕНИЕ GOOGLE_PLAY
14	GOOGLE	ANDROID WINDOWS PHONE
15	ВЕРСИЯ	МОБИЛЬНЫЙ ТЕЛЕФОН

Представления об информационной значимости наборов терминов для построения СЕИТ, степени их важности для отражения смысла научного текста были

подтверждены в ходе экспериментов с информантами. Так, для всех текстов были проведены эксперименты со стандартной инструкцией [Ягунова, 2010].

Выявление ассоциативных связей

Рассматриваемые в предложенной модели СЕИТ связи являются направленными и могут рассматриваться как отношения «общее-частное» при построении общей онтологии. Вместе с тем, построенная сеть СЕИТ может рассматриваться как основа для формирования других связей между ее узлами. Если обозначить матрицу инцидентности СЕИТ буквой A, то матрицы AA^T и A^TA будут отражать связи вхождения таких типов: если два термина-узла данной сети a_i и a_j порождают третий термин a_k, то будем считать, что такие термины связаны ассоциативной связью, назовем ее ассоциативной связью первого рода (рис. 46 а, рис. 47); если два термина-узла данной сети a_i и a_j порождаются третим термином a_k, который также входит в данную сеть, то будем считать, что такие термины связаны ассоциативной связью второго рода (рис. 46 б).



а *б*

Рис. 46 – Ассоциативные связь построенные по СЕИТ со связями, обозначенными стрелками: а) первого рода «~»; б) второго рода «≈»

На рис. 47 и 48 приведены фрагменты сети СЕИТ. На рис. 47, например, жирными кривыми обозначены ассоциативные связи между словами «мобильное» и «приложение», что объясняется наличием общего узла-термина «мобильное приложение».

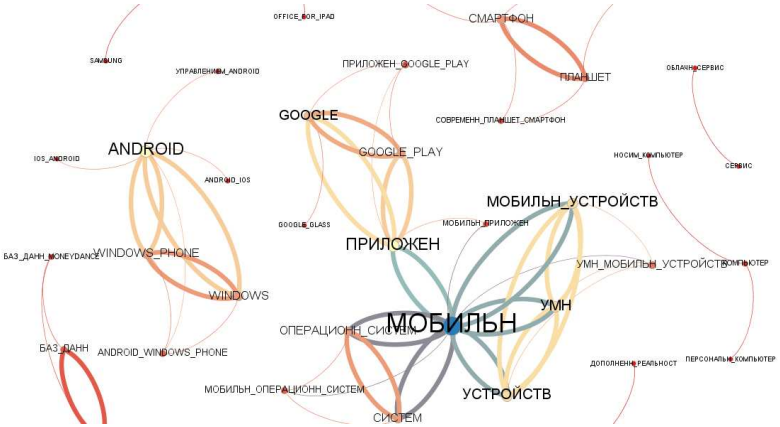


Рис. 47. Фрагмент СЕИТ размером 30+30+30 с ассоциативными связями 1-го рода

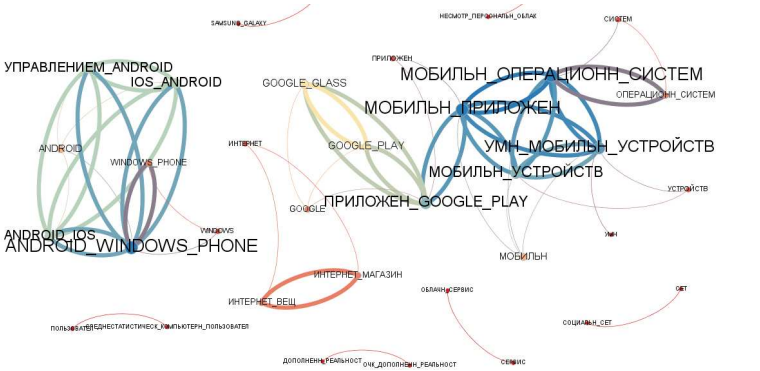


Рис. 48 – Фрагмент СЕИТ размером 30+30+30 с ассоциативными связями 2-го рода

Таким образом, с помощью алгоритма СЕИТ на основе анализа текстовых корпусов обеспечивается выбор важнейших терминов. Сеть языка, автоматически построенную с помощью предложенного алгоритма с использованием относительно небольшого тематического текстового корпуса, можно использовать в качестве основы для построения онтологии предметной области. Кроме того, СЕИТ можно использовать на практике в качестве готового к

применению средства навигации в информационных массивах, а также для организации контекстных подсказок пользователям информационно-поисковых систем.

3.9. Дублирование, сходство, упорядочивание документов

Проблема сравнительного анализа электронных текстов возникла практически одновременно с появлением возможностей обработки текстов на компьютерной технике. Вторая половина XX-го века характеризовалась становлением этого направления, бурным развитием формальной лингвистики как области дискретной математики.

Вместе с тем, следует отметить, что фундаментальные работы в области сравнения текстов проводились значительно раньше [Морозов, 1915].

В настоящее время с развитием сети Интернет задачи сравнительного анализа электронных текстов имеют отношение к ряду таких технологических направлений, как информационный поиск, обобщение и группировка информации. Развитие сети Интернет также вызвало значительный рост дублированной и заимствованной информации в различных сферах: образования, средствах массовой информации, иногда в науке [Шарапов, 2011] .

В настоящее время определены следующие пять направлений сравнительного анализа электронных текстов [Osipovs, 2009]:

1) анализ уникальности документов поисковыми машинами в интернете для предотвращения излишней индексации одинаковых документов;

2) выявление перепечатки, в том числе исходных текстов программ, случаев плагиата;

3) архивирование документов (уменьшение объемов пространства на компьютерных носителях за счет отказа от хранения подобных документов);

4) кластеризация документов по мере их сходства, выявление кластеров близких по содержанию документов, выявление основных тематик в информационных потоках;

5) поиск и фильтрация спама.

3.9.1. Исследование содержательного дублирования документов

Одной из главных особенностей потоков документальной информации является наличие большого количества документов, дублирующих друг друга. Так, о событии мирового значения напишут все средства массовой информации, причем, скорее всего, на одной из первых страниц. Потребитель же (за исключением некоторых специфических направлений аналитических исследований информационного пространства) желает получать по каждому событию одно сообщение.

Поэтому исследование характера и свойств дублирования информации приобретает в современных технологиях исключительно важное значение. В частности, очень актуальной становится задача отбора наиболее оригинальных источников, которые позволяют (по крайней мере, статистически) исключить не только формальное, но и содержательное дублирование информации.

Дублирование документов в информационных потоках зависит от разных причин, потому проведенные измерения для ранжированого по количеству публикаций списка источников показывают разный уровень, при этом информация не носит наглядного характера. В то же время, исследования свидетельствуют о стойкой тенденции: чем более производительный источник информации, тем более оно содержит заимствований из других источников.

Если некоторой документ полностью совпадает по тексту с другим документом, то говорят что имеет место четкий дубликат. Если документ по тексту совпадает не

в полной мере, но есть совпадение по содержанию, то говорят о "почти дубликате" или "нечетком дубликате".

Преодоление использования явно дублирующейся информации не представляет проблем, однако подобные по содержанию документы находятся не так легко. На практике явные дубликаты выявляются с помощью так называемых сигнатурных механизмов: контрольных сумм, хешей, но этот подход не разрешает всех проблем пользователей, для которых чаще всего не суть важно, с чем они имеют дело, с прямой перепечаткой или с перефразированием.

Важным требованием при реализации алгоритмов выявления нечетких дубликатов документов является их стойкость по отношению к небольшим изменениям содержания документов, которые анализируются. При этом следует заметить, что в технологической цепочке общей схемы выявления дубликатов документов (рис. 49) применяется вычисление контрольных сумм, хешей (MD5, SHA-1, CRC-32 и т.п.), которые были сначала разработаны для задач практической криптографии.

Необходимо отметить, что соответствующие алгоритмы ориентированы на прямо противоположную задачу, а именно, рассеивание информации, то есть небольшие изменения в содержании исходных документов должны приводить к кардинальным изменениям в контрольных суммах, кашах. Именно это противоречие является одной из предпосылок возникновения ошибок, которые встречаются в современных системах выявления нечетких дубликатов.

В рамках данной работы не будут отдельно обсуждаться вопросы точного поиска, что фактически основан на операции сравнения двух символов. Достаточно детально они обсуждаются в работах, которые уже стали классическими, например [Кормен, 2006], [Гасфилд, 2003].

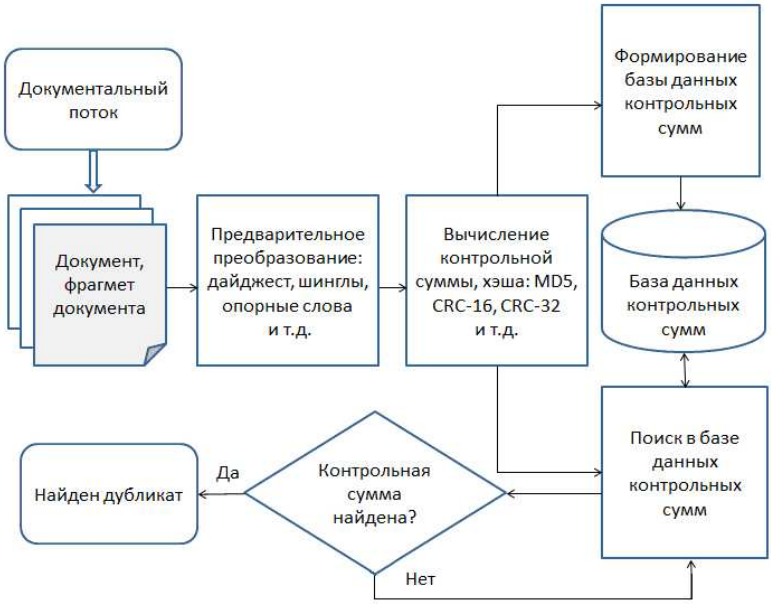


Рис. 49 – Общая схема технологии выявления нечетких дубликатов

В качестве наиболее употребляемых современных алгоритмов сравнения строчных данных рассмотрим метод Карпа-Рабина [Karp, 1987] и сигнатурные методы, в том числе метод шинглов, предложенный А. Брьодером [Broder, 2000], и его развитие – метод супершинглов. Большое практическое значение имеют методы лексических сигнатур, которые учитывают языковую природу документов, которые сравниваются. Также будут рассматриваться некоторые практические реализации методов [Зеленков, 2007].

Исследовались методы, основанные на учете повторений цепочек слов, например, метод "шинглов", обстоятельно описанный в работах [Manber, 1994], [Broder, 1997] и [Broder, 2000-1]. Однако этот эффективный во многих случаях метод поиска "почти дублей" оказался не очень чувствительным для небольших текстов с возможным перефразированием.

Естественным путем развития исследований стало обращение к статистическим подходам. Еще в 2002 году представители Яндекса опубликовали свою методику выявления дубликатов, основанную на анализе N наиболее "качественных" слов. При этом качество слов определялось экспертами, а соответствующий математический аппарат получил название "нечеткой цифровой сигнатуры". При этом используется так называемый "наивный подход" (умножения вероятностей зависимых событий – слов в сообщениях), а также элементы "ручного" отбора значимых слов (очевидно, важность отдельных слов может изменяться во времени).

Первым шагом формализации отношения подобия (условно будем считать – семантической близости, но, понятно, что второе понятие – более общее) является введение соответствующей функции $F(X,Y)$, которая ставит в соответствие некоторой паре документов (X,Y) некоторое действительное число. Функцию $F(X,Y)$ определим в окрестности [0, 1], т.е. $0 \le F(X,Y) \le 1$. Необходимым и достаточным условием совпадения X и Y является $F(X,Y)=1$.

Важное качество отношения подобия – несимметричность, то есть, в общем случае:

$$F(X,Y) = F(Y,X).$$

В дальнейшем изложении перейдем к сокращенным обозначениям, а именно $F(X,Y)>0$ обозначим как $X \prec Y$ ("$\prec$" – отношение подобия), а $F(X,Y)=1$ обозначим как $X \equiv Y$ ("$\equiv$" – отношение точного дублирования).

Очевидно, что для отношений подобия и точного дублирования справедливы правила рефлексивности:

$$A \prec A, \quad A \equiv A,$$

где A – произвольный документ.

Отношение подобия не имеет свойства симметричности. Из подобия документа *A* документу *B* не следует обратное, т.е.:

$$A \prec B \not\Rightarrow B \prec A.$$

Также не выполняется условие транзитивности:

$$A \prec B,\ \ B \prec C \not\Rightarrow A \prec C\ .$$

Действительно, например отдельный документ может быть подобный тексту из подборки, которая его включает, но сама подборка может не быть подобной к этому документа. Или документ может быть подобный до двух документов, из которых он скомпилирован, но сами оригиналы могут существенно отличаться.

Для отношения дублирования, напротив, симметричность и транзитивность выполняются:

$$A \equiv B\ \Rightarrow B \equiv A,$$

$$A \equiv B,\ \ B \equiv C\ \Rightarrow A \equiv C.$$

Заметим, что отношение, которое имеет свойства рефлексивности, симметричности и транзитивности является отношением эквивалентности [Шнейдер, 1971], в нашем случае, отношением содержательного совпадения или дублирования.

Как было отмечено ранее, свойство дублирования документов является более жестким критерием подобия, например, совпадение 3, 4 или 5 термов свидетельствуют относительно некоторой содержательной близости, то есть можно записать:

$$\text{"}\prec\text{"}\ \Rightarrow\ \text{"}\equiv\text{"}.$$

Прежде, чем непосредственно перейти к рассмотрению алгоритма Карпа-Рабина, введем некоторые обозначения и приведем один из точных алгоритмов как одну из методологических основ. Очевидно, любая строка в памяти компьютера представляется последовательностью байтов, каждый из

которых является последовательностью битов, то есть двоичных значений.

Обозначим T – двоичную строку длиной $|T| = m$; P – двоичный шаблон для поиска длиной $|P| = n$. Введем функцию:

$$H(P) = \sum_{i=0}^{n} 2^{n-i} P(i),$$

где $P(i)$ – i-й бит строки P.

Также определяется функция от подстроки T длиной n (T_r^n), которая начинается с r–го бита:

$$H(T_r^n) = \sum_{i=0}^{n} 2^{r+n-i} T_r^n(r+i-1).$$

Из того, что любое целое число можно единственным способом представить в виде суммы положительных степеней двойки, следует, что подстрока T_r^n входит в T начиная с з позиции r в том и только в том случае, когда $H(P) = H(T_r^n)$.

При поиске подстроки в строке, таким образом, при сравнении $H(P)$ и $H(T_r^n)$, из того, что представление 2^n числа требует n битов, следует, что необходимые для сравнения числа экспоненциально велики, т.е. задача сравнения оказывается экспоненциально сложной.

В 1987 г. Р. Карп и М. Рабин обнародовали алгоритм, который получил название метода рандомизированных дактилограмм, в котором существенно снижена сложность вычислений. Но этот алгоритм позволяет утверждать относительно вхождения подстроки в строку не абсолютно точно, а с некоторой высокой вероятностью. Предоставим сжатое содержание этого алгоритма.

Пусть $H_p(P) = H(P) \mod p$ – остаток от деления $H(P)$ на p. В случае, если p – простое число ($p << n$), остаток от деления $H_p(P)$ и $H_p(T_r^n)$ на p называют дактилограммами P и T_r^n по модулю p. Известно, $0 \le H_p(P),\ H_p(T_r^n) \le p-1$.

Если P входит в T начиная с позиции r, то $H_p(P) = H_p.(T_r^n)$, но обратное не верно. Говорят, что когда $H_p(P) = H_p.(T_r^n)$, но P не входит в T, то имеет место ошибочное совпадение P и T_r^n с позиции r. Для оценки вероятности отсутствия ошибочных совпадений введем обозначение $\pi(q)$ – количество простых чисел, не превышающих q. Доказано, что имеет место следующее неравенство:

$$\frac{q}{\ln q} \le \pi(q) \le 1.26\frac{q}{\ln q}.$$

Кроме того, имеет место теорема, справедливая для P и T, при условии $nm \ge 29$ ($n = |P|$, $m = |T|$). Если I – любое положительное число, а p – случайным образом выбранное простое число, которое не превышает I, то вероятность ошибочного совпадения P и T не превышает $\pi(nm) / \pi(I)$.

Отсюда следует такой алгоритм случайной дактилограммы для поиска вхождений P в T:

1. Выбрать позитивное целое число I.

2. Случайным образом выбрать простое число p, которое не превышает I, и вычислить $H_p(P)$.

3. Для каждой позиции r в p вычислить $H_p(T_r^n)$ и сопоставить с $H_p(P)$. Если они равны, то или объявить

о вероятном совпадении, или в явном виде проверить совпадение P с T, начиная с позиции r.

Так как каждое $H_p(T_r^n)$ можно подсчитать за определенное время из $H_p(T_{r-1}^n)$, то алгоритм дактилограммы реализуется за время $O(m)$, исключая время явной проверки объявленного совпадения.

Иногда, ввиду слишком низкой вероятности ошибки, оказывается излишним требование явной проверки совпадения.

Известно, что если $I = nm^2$, то вероятность ошибочного совпадения не превышает $2{,}53/m$. Действительно:

$$\frac{\pi(nm)}{\pi(nm^2)} \le 1{,}26\frac{nm}{nm^2}\frac{\ln(nm^2)}{\ln(nm)} = 1{,}26\frac{1}{m}\left(\frac{\ln n + 2\ln m}{\ln n + \ln m}\right) \le \frac{2{,}53}{m}.$$

Например, если $n = 250$, $m = 4000$ и, соответственно, $I = 4\times10^9$. Тогда вероятность того, что хотя бы одно из выявленных совпадений будет ошибочным меньше чем $2{,}53/4000$, т. е. не превысит 0,001.

Через десять лет после обнародования алгоритма рандомизированных дактилограмм Кара-Рабина А. Бродер и его соавторы [Broder, 1997] представили свой алгоритм, который базируется на оценке подобия документов по совпадению последовательностей из определенного количества соседних слов. Такие последовательности авторы назвали шинглами (от англ. Shingles – "чешуйки", "опоясывающий лишай"). Необходимо отметить, что разные шинглы формируются из последовательностей слов внахлест, а не впритык, встык, т.е. следующий шингл начинается со следующего слова, а не из слова с номером, больше на длину шингла. Два документа считаются дубликатами, если множества их шинглов существенно пересекаются.

Таким образом, разбивая текст на последовательности слов (их еще называют N-граммами), мы получаем набор шинглов в количестве $N-n+1$, где N – количество слов в документе.

При этом на шинглы разбивается каждый экземпляр документов, которые сравниваются. Поскольку количество шинглов, соответствующих каждому документу, является таким образом, достаточно большим, было предложено несколько методов их уменьшения при получении репрезентативных подмножеств для сравнения.

Первый предложенный метод заключался в том, что рассматривались лишь те шинглы, чьи дактилограммы, вычисленные по алгоритму Карпа-Рабина, делятся без остатка на некоторое число m. Основной недостаток этого подхода – зависимость выборки от длины документа и потому небольшим по размеру документам отвечают очень короткие выборки, что приводит к низкому качеству выявления дубликатов.

В соответствии с другим предложенным методом отбиралось лишь фиксированное количество s шинглов с наименьшими значениями дактилограмм или оставались все шинглы, если их общее количество не превышает s.

Дальнейшим развитием метода шинглов являются методы "супершинглов" и "мегашинглов" [Broder, 2000-1].

Метод супершинглов заключается в том, что для каждого документа выбираются случайные наборы шинглов в количестве 84. Для каждого выбранного шингла подсчитывается дактилограмма Карпа-Рабина. После этого 84 шинглов разбиваются на 6 групп (супершинглов) по 14 шинглов в каждой. В итоге каждый документ представляется 6 супершинглами. Оказывается, что при условии подобия документов на уровне 95%, вероятность совпадения 2-х супершинглов представляет приблизительно 90%, а если подобие

между документами лишь 80%, то вероятность совпадения по меньшей мере двух шинглов представляет лишь 2,6%. Таким образом, для эффективного сравнения документов оказывается достаточным исследовать совпадение лишь одной пары шинглов.

Идея объединения шинглов получила дальнейшее развитие в методе мегашинглов, который заключается в том, что для каждого документа рассматриваются все пары его супершинглов. Количество таких пар составляет $C_6^2 = \frac{6 \cdot 5}{2} = 15$. Из вышеприведенного можно сделать вывод: два документа являются по меньшей мере нечеткими дубликатами, если у них совпадает хотя бы один мегашингл.

Необходимо отметить, что методы, связанные с подсчетом шинглов, не являются эффективными при сравнении небольших по размеру документов.

Поскольку сравнение документов может учитывать языковую природу последних, что не используется в двух приведенных выше формальных алгоритмах, учет этой особенности в некоторых случаях повышает эффективность обнаружения нечетких дубликатов. В этих случаях, в отличие от алгоритма шинглов, в качестве основных единиц измерения используются слова из документов. Этот класс алгоритмов предполагает фокусирование на семантическом подобии (сходстве) документов, не уделяя внимания анализу их структуры. Наиболее часто используется построение словарей опорных слов, осуществляется сравнение словарей отдельных документов между собой.

Существует несколько подходов для получения численных значений степени близости между документами. Если обозначить множество опорных слов документа как X, а $|S(X)|$ – как меру появления слов в документе, которая определяется как количество

опорных слов, то меры близости рассчитываются следующим образом:

Косинус:

$$simCos(A,B) = \frac{|S(A) \cap S(B)|}{\sqrt{|S(A)|^2 + |S(B)|^2}};$$

Коэффициент Жаккарда:

$$simJaccard(A,B) = \frac{|S(A) \cap S(B)|}{|S(A) \cup S(B)|};$$

Коэффициент Дайса:

$$simDais(A,B) = \frac{|S(A) \cap S(B)|}{|S(A)| + |S(B)|}.$$

В работе [*Антонова, 2011*] показано, что в некоторых случаях эффективною является несимметричная упрощенная мера близости:

$$simNSL(A,B) = \frac{|S(A) \cap S(B)|}{|S(A)|},$$

а также симметричная, рассчитываемая как сумма двух несимметричных:

$$simSSL(A,B) = \frac{|S(A) \cap S(B)|}{|S(A)|} + \frac{|S(A) \cap S(B)|}{|S(B)|}.$$

Один из подходов, основанный на расчете лексической дактилограммы, называется *I-Match* и был предложен в работе [Chowdhury, 2002]. Для этого в исходном множестве (коллекции) документов строится словарь, который включает слова со средними значениями инверсной частоты документа *IDF*.

Значение *IDF* рассчитывается по формуле:

$$IDF = \log \frac{|D|}{|t_j \in d_j|},$$

где $|D|$ – количество документов в коллекции; $|t_j \in d_j|$ – количество документов, где встречается t_j.

Слова со средними значениями *IDF,* как правило, обеспечивают более точные результаты при выявлении нечетких дубликатов. После этого для каждого документа выбираются слова, которые также входят в сформированный словарь. Множество этих слов упорядочивается и рассчитывается соответствующая контрольная сумма (применяется хэш-функция *SHA-1*), которая и называется *I-Match.* Два документа считаются нечеткими дубликатами, если в них соответствующие значения *I-Match* совпадают.

Приведенный алгоритм с точки зрения вычислительной сложности более эффективен, чем алгоритм шинглов, кроме того он может применяться для сравнения небольших по размеру документов. Однако применение жесткой хэш-функции *SHA-1* делает его неустойчивым к небольшим изменениям содержания документов.

Еще один сигнатурный метод, применяемый в службе Яндекс.Новости для выявления сюжетов, базируется на определении «опорных» слов [Ilyinsky, 2002]. В соответствии с этим алгоритмом выбирается множество из *N* слов, которые называются опорными. После этого каждому документу ставится в соответствие *N*-мерный двоичный вектор, координаты которого соответствуют присутствию слов (принимают соответствующие значения 0 или 1). Этот двоичный вектор и называется сингантурою документа. Два документа считаются подобными, если их сигнатуры совпадают.

Принципы выбора опорных слов могут различаться в разных реализациях. В частности,

принцип выбора опорных слов в системе контент-мониторинга InfoStream приведен ниже.

Ниже описываются критерии качества обнаружения подобных документов, основанные на анализе некоторых свойств так называемой матрицы подобия, как симметричность и транзитивность [Ландэ, 2008], рассматриваются аналитические выражения для расчета этих критериев, а также приводятся результаты экспериментов на многоязычных текстовых корпусах, которые формируются с помощью системы контент-мониторинга.

На практике каждому документу D_i из контрольного документального корпуса по алгоритму совпадения термов в сигнатурах (в разных экспериментах менялось необходимое количество совпадающих термов) ставился в соответствие вектор с элементами:

$$a_{ij} = \begin{cases} 1, & D_i \equiv D_j, \\ 0, & D_i \not\equiv D_j. \end{cases}$$

Условие симметричности в этих обозначениях записывается следующим образом:

$$\forall\ i,\ j:\ a_{ij} = a_{ji},$$

а условие транзитивности:

$$\forall\ i,\ j,\ k:\ a_{ij} = 1,\ a_{jk} = 1 \Rightarrow a_{ik} = 1.$$

Согласно приведенным соображениям были предложены критерии, основанные на вычислении коэффициентов симметричности (S) и транзитивности (T) для матрицы подобия. На контрольном документальном корпусе, изменяя количество сравниваемых в сигнатурах термов, были получены различные значения соответствующих коэффициентов. Коэффициент симметричности рассчитывается

следующим образом:

$$S = 2\frac{\sum_{i}^{N}\sum_{j\neq i}^{N} a_{ij}a_{ji}}{\sum_{i}^{N}\sum_{j\neq i}^{N} a_{ij}},$$

а коэффициент транзитивности T определяется по формуле:

$$T = \frac{\sum_{i}^{N}\sum_{j\neq i}^{N}\sum_{\substack{k\neq i\\ k\neq j}}^{N} a_{ij}a_{jk}a_{ik}}{\sum_{i}^{N}\sum_{j\neq i}^{N}\sum_{\substack{k\neq i\\ k\neq j}}^{N} a_{ij}a_{jk}}.$$

где N - количество документов в контрольном корпусе.

Очевидно, что коэффициент симметричности, который рассчитывается таким образом, ассоциируется с точностью при определении дубликатов документов, а уровень транзитивности – с полнотой.

Вместе с тем следует отметить, что проверка коэффициентов асимметрии и транзитивности может использоваться только для формальной проверки приближения отношения к свойствам эквивалентности. Само определение того, что эта эквивалентность – содержательное дублирование подтверждается экспертами-аналитиками. Приведенный выше алгоритм кроме своего эмпирического подтверждения имеет то преимущество, что позволяет варьировать некоторым числом (количеством сравниваемых термов в сигнатурах), значение которого можно подобрать с учетом оптимизации двух названных коэффициентов.

Следует выделить две проблемные области в выявлении дубликатов по представленному алгоритму. Во-первых – это некорректная во многих случаях обработка коротких текстов (сообщений), которые часто

вырождаются только в один заголовок. Выявление значимых слов в таких сообщениях – открытая проблема, актуальная, например, при реализации метапоисковых систем в веб-пространстве, где приходится иметь дело только с короткими рефератами (сниппетами) документов.

Вторая проблема связана с длинными документами, обзорами, дайджестами. Опорные слова в словесных сигнатурах таких документов могут не отражать контента каждой составляющей обобщенного документа, а иногда охватывать значительную часть словаря соответствующего языка.

Наряду с вышесказанным, необходимо заметить, что устранение дублированных сообщений в информационных потоках нужно далеко не всегда. Существует ряд задач, в которых используется факт дублирования текстов сообщений в различных источниках, например, при определении важности сообщения (сообщение многократно дублируется на сайтах и в СМИ) или при определении эффективности PR-кампаний (подсчет републицируемых пресс-релизов и т.д.).

Одним из ключевых аспектов развития современных информационных технологий является специфика взаимоотношений между информационными агентствами (ИА), которые традиционно играют роль поставщиков информации, и СМИ, которые являются основным ее потребителем. Видимо, эти отношения в значительной степени устарели и требуют серьезных корректив как в технологическом плане, так и в плане организационном, включая законодательное регулирование.

Главная причина такого положения дел заключается в быстром расширении влияния на информационные процессы сетевых технологий и, разумеется, в первую очередь Интернет. Развитие этих технологий привело к качественным изменениям в структуре всего процесса информирования общественности на всех его звеньях, в результате чего

ситуация требует кардинального пересмотра основных механизмов, лежащих в основе функционирования медийных средств.

Информационные агентства поставляют своим подписчикам информацию на условиях, которые на сегодняшний день выглядят по меньшей мере странно. В частности, типичным условием по использованию материалов ИА является запрет на размножение и распространение их любыми способами. Таким образом агентства стараются защитить свою продукцию от копирования, часто ссылаясь на законодательство об авторских правах. Вместе с тем в статье 10 Закона Украины «Об авторском праве и смежных правах» также предусмотрено, что сообщение о новости или текущие события не охраняются авторским правом.

В аналогичном законе Российской Федерации в статье 8 говорится, что «сообщения о событиях и фактах, имеющие информационный характер» не охраняются авторским правом. Таким образом, условия, декларируемые большинством ИА со ссылкой на законодательство об авторских правах, являются неправомерными, по крайней мере, по отношению к их основной продукции – информационных сообщений.

Не лучше ситуация и с содержательным аспектом проблемы. Никто, конечно, не ставит под сомнение авторские права на те материалы, которые действительно имеют авторы в обычном смысле слова (интервью, аналитические разработки, эксклюзивные репортажи и т. д.). Но говорить об авторских правах на сообщение об официальном визите главы государства или вступления в силу нового закона явно лишено конкретного смысла. Мы уже не говорим о текстах законов, указов и т. п., для которых законодательно предусмотрен порядок обнародования.

Как всегда в подобных ситуациях, новые тенденции начинают прокладывать себе дорогу, не дожидаясь официальных решений, что неизбежно приводит к перераспределению не только ресурсов, но и функциональных ролей участников коммуникации.

Поэтому для выработки обоснованных рекомендаций желательно было бы сначала разобраться в том, что и как происходит на самом деле.

В связи с этим как научный, так и практический интерес вызывает вопрос, в какой мере материалы, доступные платным подписчикам основных ИА, становятся доступными в открытом доступе на информационных сайтах. Ценность информационных сообщений во многом определяется оперативностью, поэтому отдельной задачей является оценка запаздывания публикаций в Интернет по сравнению с временем рассылки соответствующих сообщений. Забегая вперед, скажем, что почти в третьей части рассмотренных случаев время задержки оказалось отрицательным, т.е. ИА копировали сообщения с веб-сайтов со значительным опозданием.

При проведении исследований авторы получили возможность доступа к подписным материалам ведущих ИА, представленных в украинском информационном пространстве. Кроме того, в распоряжении авторов находилась система контент-мониторинга InfoStream, с помощью которой в реальном масштабе времени сканируется более 5000 информационных сайтов, представленных в украинском и российском сегментах веб-пространства. Таким образом, в ходе исследования рассматривались два текстовых корпуса (точнее, наборы «словесных сигнатур» текстов [Ландэ, 2006-1], представленных в этих корпусах) – сообщений ИА и текстов, отсканированных с веб-пространства.

В качестве примера рассматривались сообщения ИА по общеполитической тематике, поступившие в течение 5-25 ноября 2007 года. Их объем составил 8955 документов. Эти сообщения сравнивались с текстами, отсканированными из Интернета в течение всего ноября 2007 года, количество которых составило более 1 млн. документов.

Технически задача нахождения дубликатов (в данном случае речь идет именно о дубликатах, а не сообщениях по той же тематике, но другими словами,

если учитывались перепечатки с незначительными искажениями) решалась методом, который описан в работе [Ландэ, 2006-1]. Этот метод относится к группе методов нахождения «нечетких» дубликатов [Никконен, 2007], [Bourdaillet, 2007], основанных на выделении некоторого множества опорных слов, имеющих наибольшие значения *TF IDF*.

Анализ новостных сообщений показал, что при перепечатке материалов чаще всего остаются без изменений несколько первых предложений текста или первый абзац. Многие из недобросовестных изданий перепечатывают содержание сообщений, просто меняя названия (работа хедлайнеров). И этот фактор был также учтен. Такой вид дублирования успешно преодолевается путем использования сигнатур: шинглов, хешей и т.д. (но уже без учета заголовков).

В качестве некоторых «инвариантов» для отдельных сообщений использовались лингвистические сигнатуры: цепочки из 12 опорных слов, прошедших процедуру морфологической обработки (стемминга). Такое небольшое количество термов в цепочке, которая является своеобразной словесной сигнатурой, объясняется небольшой средней длиной новостных сообщений (2000-3000 символов).

В результате проведенных исследований удалось получить следующие данные:

- из 8955 сообщений ИА на сайтах было опубликовано 5567 сообщений (62%);
- общее количество перепечаток на разных сайтах составило 39901 (456%).
- Соответствующее гиперболическое распределение приведено на рис. 50;
- количество перепечаток с положительным временем запаздывания (из материалов ИА на сайты) составило 28933 (73%);
- количество перепечаток с отрицательным

временем запаздывания (перепечаток с Интернет ленты ИА) составило 10968 (27%).

Ранжированный график распределения сообщений ИА по времени задержки публикаций приведен на рис. 51, на котором четко видны экстремальные отклонения в начальной и конечной области. Отклонение в начальной области характеризует большое время задержки включения в ленты ИА материалов, размещенных, как правило, на сайтах органов государственной власти (инертность ИА, отсутствие у них средств для мониторинга веб-пространства).

Отклонения в конечной области объясняются задержками перепечаток на веб-сайтах, которые получили со временем некоторое новое продолжение.

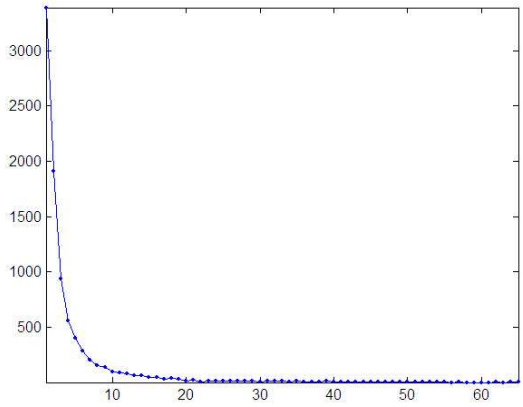


Рис. 50 – Количество сообщений ИА (ось ординат), ранжированных по количеству перепечаток на сайтах (ось абсцисс)

Вместе с тем центральная область графика (от 1000-го по 5000-е сообщения) имеет стабильный характер со средним значением около получаса.

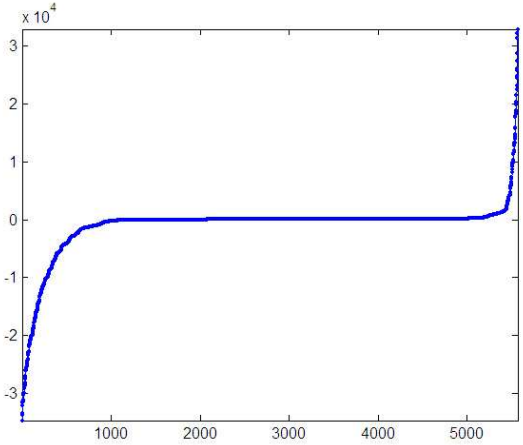


Рис. 51 – Распределение сообщений ИА (ось абсцисс) по времени запаздывания в минутах (ось ординат)

Отклонения в конечной области объясняются задержками перепечаток на веб-сайтах сообщений, которые получили со временем некоторое новое продолжение. Вместе с тем центральная область графика (от 1000-го по 5000-е сообщения) имеет стабильный характер со средним значением около получаса.

Массовый характер перепечаток позволяет делать выводы, что все сообщения, интересные администраторам соответствующих сайтов, были перепечатаны. Видимо, примерно 38% сообщений ИА оказались им недостаточно интересными.

В результате исследования оказалось, что методы определения нечетких дубликатов сообщений, развитые в последние годы, оказались полезными в данном применении. Результаты заставляют задуматься, за что платят подписчики информационным агентствам сегодня, когда большая часть информации с минимальной задержкой доступна в Интернет, а полноту могут обеспечить системы контент-мониторинга? Видимо, за аналитический подбор этой информации, репрезентативность и достоверность. То

есть, информационное агентство, если оно желает выжить в современных условиях, должно уделять повышенное внимание именно аналитической обработке информации, превращаясь в агентство информационно-аналитическое.

3.9.2. Семантическое сходство документов

На практике при поиске новостной информации всегда возникает задача выявления информационных кластеров (сюжетов), состоящих из отдельных документов, и их ранжирования по некоторым признакам, которые должны обеспечить не только выявление важнейшей темы, но и «веерное» многоаспектное освещение всех наиболее значимых аспектов информационных сюжетов. Эта задача решается во многих системах с использованием различных подходов и алгоритмов. При этом неизменным остается технологическая цепочка: построение семантической сети из информационных сообщений, кластеризация – выявление наиболее взаимосвязанных групп, т.е. информационных сюжетов, «взвешивания» (оценка важности, актуальности) и наглядная визуализация значимых из них.

При выделении сюжетных цепочек для определения попарной текстуальной близости отдельных документов, как правило, используются алгоритмы обнаружения подобных документов, ставшие уже традиционными в поисковых системах. Так матрица попарной близости документов обрабатывается алгоритмами кластеризации, такими как LSA/LSI, k-means, суффиксних деревьев и так далее. Выделенные классы документов и являются информационными сюжетами.

Для предъявления пользователям информационные сюжеты должны быть ранжированы. Основные факторы, влияющие на ранжирование по важности – оперативность информации и размер сюжетной цепочки. Под оперативностью понимается некоторая функция от времени публикации всех

документов в информационном сюжете, а размер отражает общий интерес к конкретной теме. Во всех этих подходах центральная задача состоит в отождествлении документов, относящихся к одному сюжету и выделения непересекающихся сюжетов.

На рис. 52 представлена типовая схема выявления информационных сюжетов.

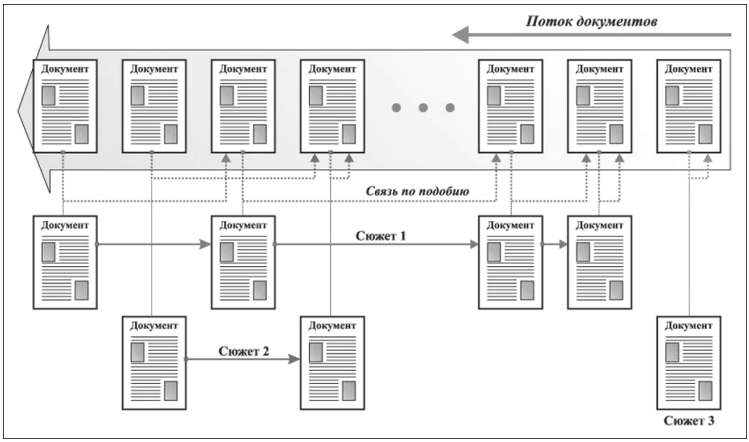


Рис. 52 – Типовая схема выявления информационных сюжетов

Последнее по времени генерации информационное сообщение (документ) сравнивается с предыдущими, оценивается уровень их подобия. Если уровень сходства превышает некоторый порог, то рассматриваемый документ считается принадлежащим информационному сюжету, к которому относится ранее сгенерированный документ. Если подобных документов не найдено, то фиксируется новый сюжет, состоящий на данный момент из одного документа.

Для результирующего отображения каждого отдельно взятого инфосюжета используются отобранные по текстуальной близости документы из различных источников, отсортированные в хронологическом порядке.

При этом сюжеты могут быть дайджестами,

интегрирующими документы по теме, а также содержать уникальную информацию от отдельных документов, то есть реферирование сюжета в этом случае сводится не к свертыванию информации, а к построению расширенной версии, по сравнению с любым документом по сюжетной цепочки.

Например, в системе Яндекс.Новости (http://news.yandex.ru) для выделения информационного сюжета строится матрица попарной близости документов, которая обрабатывается алгоритмом кластеризации с эмпирически подобранными параметрами (в частности, радиусом метрики близости).

Для того, чтобы увеличить связность крупных сюжетов дополнительно используется кластеризация второго уровня, обеспечивающая сбор атомарных кластеров в большие.

Все сообщения о результатах поиска на сайте сгруппированы, при этом ранжирование инфосюжетов построено на стандартных для Яндекса принципах ранжирования выдачи. Оно основано на числе и ранге новостей внутри новостных сюжетов, при этом ранг одной новости определяется как ее "свежесть" с учетом приоритетов удовлетворения критериев поиска.

В системе *InfoStream* (http://infostream.ua) тематическая близость документов определяется на основе нормированных последовательностей наиболее весомых терминов, входящих в каждый документ [Григорьев, 2005].

Последовательности подобных (с определенным коэффициентом взаимной близости, который превышает некоторый установленный эмпирически уровень) документов образуют цепочки. При этом каждый документ попадает в какую либо цепочку, в крайнем случае, состоящий только из него самого. Затем цепочки взвешиваются по длине и оперативности, после чего пользователю предъявляется определенное количество наиболее важных тематических

инфосюжетов.

Для репрезентации сюжетной цепочки заголовки документов также взвешиваются относительно ключевых слов, соответствующих сюжету, а затем со всех заголовков выбираются наиболее "весомые" для отображения **(**рис. 53).

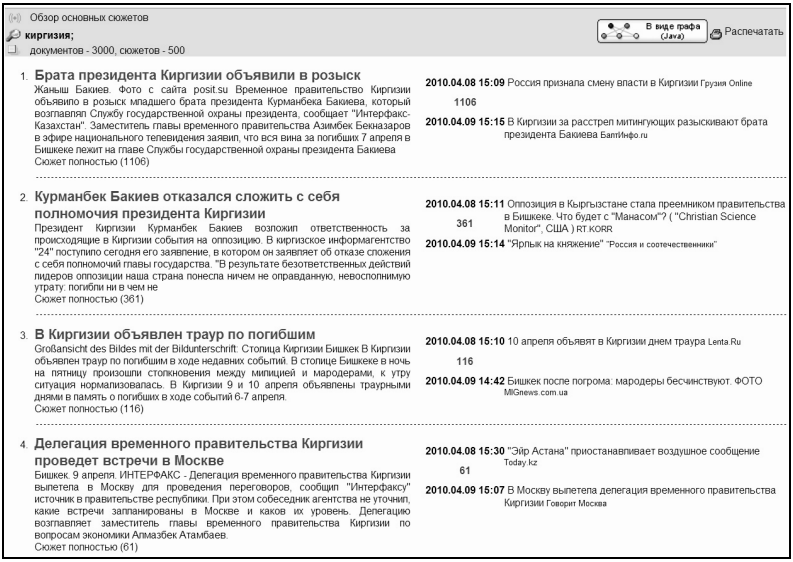


Рис. 53 – Пример отображения инфосюжетов в системе InfoStream

Ниже описан предложенный в [*Ланде, 2008*] подход, с помощью которого можно обнаружить нечеткие дубликаты документов, приведенных на разных языках, а именно, на украинском и русском. Рассматриваемая процедура обнаружения дубликатов построена на использовании методов извлечения опорных слов на основе эмпирико-статистических свойств текста с помощью частотного морфологического словаря, а также перевода этих слов на другой язык.

Для анализа и оценки результатов, излагаемых ниже, использовался достаточно мощный информационный ресурс – ретроспективная база

данных системы контент-мониторинга InfoStream [*Григорьев, 2005*], которая применяется для решения задач автоматизированного сбора новостной информации из открытых веб-сайтов и обеспечения доступа к ней в поисковых режимах. Эта разработанная в компании ElVisti система в настоящее время охватывает более 5000 источников, а ретроспективные базы данных системы представляют собой корпус объемом более 100 млн. документов. Следует отметить, что количество дублирующихся сообщений в системе InfoStream значительно меньше, чем во всем веб-пространстве. Это объясняется подбором источников для сканирования, в число которых входят только те, которые хоть изредка публикуют оригинальные материалы. Принцип обнаружения значимых ключевых слов (далее будем называть их термами или опорными словами) в системе InfoStream базируется на законе Ципфа и сводится к выбору слов со средней частотой появления (наиболее часто встречающихся слова игнорируются с помощью «стоп-словаря», а редкие слова из текстов сообщений не учитываются).

Выявление дублирующихся по содержанию новостных сообщений в системе InfoStream выполняется методом, заключающимся в выявлении в различных документах общих термов, цепочки которых образуют лингвистические сигнатуры сообщений. Определение дублей документов заключается в выявлении опорных слов с использованием морфологических частотных словарей, в которые вошли существительные и общеизвестные фамилии и названия фирм и организаций. Расчет коэффициентов производится на основании весового подхода – модификации стандартного подхода *TF IDF*, *Okapi BM25*.

Полученные результаты позволили вплотную подойти к решению проблемы эффективного автоматизированного обнаружения плагиата в текстах небольших объемов. Эта проблема сегодня имеет большой резонанс, но существующие методы ее решения часто не раскрываются из-за опасения обесценивания наработанных механизмов.

Во многих случаях в основу поиска нечетких дубликатов, в частности плагиата, положено сравнение текстов [Ашура, 2006], [Stone, 2003]. Изначально проводится сравнение текстов в целом, а дальше происходит разбивка на абзацы и затем поиск конкретных фрагментов текста в других документах. В других же случаях используется поиск по словам или словосочетаниям. Учитывая количество терминов с малой частотой, найденных при проверке, как результат получают подтверждение того, что документ или фрагмент является плагиатом, или наоборот. Многочисленные системы выполняют поиск не только фрагментов, но даже проводят достаточно сложный анализ текста, включающий использование такого метода, как метод Флэша – который позволяет вычислить индекс «легкости» текста [Нейл, 2005].

Анализируя показатели индексов абзацев в проверяемой работе, можно идентифицировать наиболее вероятные абзацы плагиата, поиск которых затем проводится по внутренней базе данных, или в веб-пространстве.

Для выделения ключевых слов в тексте в рамках рассматриваемого подхода используется векторная модель документа, в соответствии с которой каждому слову документа приписывается его весовой коэффициент. Чем больше вес слова, тем больше это слово характеризует документ. В рамках этой работы было проверено два подхода к вычислению весовых коэффициентов слов.

Мера TF IDF часто применяется в задачах информационного поиска, например, как один из критериев релевантности документа поисковому запросу, для расчета степени близости при формировании документальных кластеров:

$$TF = \frac{n_i}{\sum_k n_k},$$

где n_k – количество употреблений слова, *IDF* (inverse document frequency – обратная частота документа) –

величина, обратная частоте, с которой некоторое слово встречается в документах массива.

IDF уменьшает вес часто употребляемых слов:

$$IDF = \frac{|D|}{N_i},$$

где $|D|$ – количество документов в массиве; N_i – количество документов, в которых встречается слово.

Таким образом, мера *TF IDF* является произведением двух сомножителей: *TF* и *IDF*. Известно, что стандартный алгоритм *TF IDF* не совсем корректно отрабатывает на больших текстовых массивах. Как альтернативу предложено его модификацию – *Okapi BM25* [Salton, 1988]:

$$TF \cdot IDF(w_i) = IDF(w_i) \frac{f(w_i, D) \cdot (k_1 + 1)}{f(w_i, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{avgdl})},$$

где $f(w_i, D)$ – частота слова w_i в документе *D*; $|D|$ – длина документа *D* (число слов) *avgdl* – средняя длина документа в коллекции; k_1 и b – свободные параметры, обычно выбраны как k_1 = 2.0 и b = 0.75. $IDF(w_i)$ - инверсная частота документа, которая вычисляется по формуле:

$$IDF(w_i) = \log \frac{N - n(w_i) + 0.5}{n(w_i) + 0.5},$$

где N – общее число документов в массиве; $n(w_i)$ – количество документов, содержащих сроки w_i.

Суть данного подхода заключается в том, что в отличие от первичного подхода *TFI DF* в *Okapi BM25* учитывается длина документа.

Процедуру выявления дубликатов можно представить в виде нескольких этапов (рис. 54):

- создание морфологических словарей;
- создание частотных словарей – обучение системы;

- создание словарей переводов;
- построение программамы поиска ключевых слов;
- создание процедуры поиска дубликатов, приведенных на разных языках.

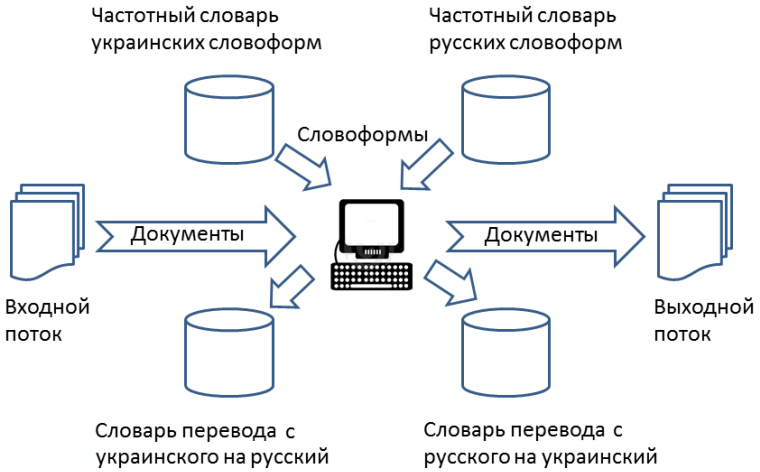


Рис. 54 – Функциональная схема поиска и перевода ключевых слов

Сначала создаются морфологические словари, которые для каждой словоформы содержат ее нормальную форму. Это нужно для того, чтобы в дальнейшем можно было привести все найденные словоформы к нормальной форме. После этого создается частотный словарь на базе морфологического словаря, в котором записывается частота каждой словоформы, найденной в процессе «обучения» частотного словаря на тестовом массиве документов.

Для построения электронных морфологических словарей была взята имеющаяся электронная версия словаря Зализняка, который насчитывает около 93 тыс. слов в нормальной форме для русского языка и бесплатный словарь ispell, который насчитывает около 1 миллиона украинских словоформ, соответственно, для украинского языка. Морфологические словари были

дополнены известными фамилиями и названиями учреждений и организаций, которых в них не было.

Для выявления опорных слов документов был построен частотный словарь, в котором для каждого слова записано количество его появлений в некотором большом массиве документов, а также количество документов, в которых нашлось слово.

Для создания частотного словаря был взят корпус документов за 2007 год, которые сканируются с Интернет системой контент-мониторинга InfoStream. Корпус состоял из текстов веб-публикаций на украинском (1344086 документов) и русском языке (2399367 документов). При машинном обучении частотного словаря с каждого документа в корпусе вытягивались словоформы, которые (с определенной вероятностной погрешностью) были приведены к нормальной форме. При этом подсчитывалось количество как словоформ, так и нормальных форм в документах, а также подсчитывалось количество документов, в которых встретилась словоформа или нормальная форма.

Для эффективности поиска опорных слов в результирующие словари входили только те слова, которые встретились в документальном корпусе более двух раз. Также было решено использовать только существительные.

Обучение словаря проходило в три этапа. Первый этап заключался в разделении документов на словоформы и записи полученных словоформ с информацией о номере документа, во временный файл. На вход программы подается документ, программа разделяет документ на словоформы, и записывает все в файл. На втором этапе созданный файл сортируется по словоформе и номеру документа. Далее подсчитывается количество вхождений одной словоформы и количество документов, в которых она встретилась. Найденные частоты записываются в частотный словарь, после чего происходит поиск нормальной формы. В новом файле сохраняется нормальная форма и номера документов.

Для решения задачи построения параллельных текстовых корпусов в результирующие словари отбираются все словоформы существительных.

Описанный подход предполагает использование алгоритма решения контекстной неоднозначности, так как омонимия является существенной проблемой при определении опорных слов документа, например, слово «села», которое в практике русского языка может быть множеством от слова «село», а также производной от глагола «садиться», может некорректно переводиться и использоваться на украинском языке, потому что слово «село» переводится на украинский как «деревня», а слово «сажается» – «садиться». Неправильный выбор нормальной формы может привести к тому, что в одинаковых по информационному содержанию документах, приведенных на разных языках, будут использованы различные опорные слова. Для решения этой проблемы может применяться, как оказалось позже, эффективный и достаточно быстрый алгоритм, что особенно важно, так как этап обучения частотных словарей и этап их использования связаны с обработкой больших объемов текстовой информации.

В табл. 5 приведен пример обучения частотного словаря для слов «сажается» и «село». Предложено правило, согласно которому, если в систему поступила словоформа, которая на практике может приводить к нескольким нормальным формам (например, для словоформы «деревни» допустимые нормальные формы «село» и «садится»), то так называемые «индексы нормальных форм» для этой словоформы увеличиваются на единицу. В примере в русскоязычном текстовом корпусе словоформа «села» встретилось 20 раз, словоформа «деревня» – 50 раз, словоформа «сели» –10 раз, а словоформа «селом» – 30 раз. В результате обучения в словари попадают слова «село» с индексом нормальной формы 100 и «садится» с индексом 80, соответственно, в дальнейшем при отборе опорных слов преимущество будет предоставляться слову «село».

Табл. 5. Пример обучения системы

Словоформа	Количество	Индекс нормальных форм
села	20	садиться → +20 село →+20
село	50	садиться → +50 село → +50
сели	10	садиться → +10
селом	30	село → +30
		село = 100 садиться = 80

В рамках данных исследований использовались словари переводов с русского языка на украинский, и с украинского на русский язык. Исходные данные для построения словарей переводов были получены путем перевода существительных в нормальной форме существующими программами перевода текстов. Если одному слову соответствовало несколько переводов, то выбиралось наиболее употребляемое значение в соответствии с частотным словарем.

Программа формирования опорных слов и их переводов загружает стоп-словари для каждого языка. Загрузка стоп-словарей проходит в два этапа. Первый этап осуществляется при старте программы, а второй – когда выбираются опорные слова документа. Первый этап происходит при загрузке морфологических словарей, при этом отсеиваются все нормальные формы, которые соответствуют данным словам в словарях, и находятся в стоп-словаре. После этого происходит загрузка словарей переводов.

Для каждого документа, который считывается из входного потока, происходит его распределение по словоформам. После этого происходит поиск нормальной формы для каждой словоформы. В случае омонимии выбирается нормальная форма, наиболее частотная в словаре. Далее происходит подсчет количества словоформ. Опорные слова извлекаются с помощью формулы *Okapi BM25.* После вычисления весовых коэффициентов происходит ранжирование

нормальных слов и выбираются первые двенадцать опорных слов.

Полученные двенадцать опорных слов переводятся с одного языка на другой с помощью словарей переводов. Все опорные слова и слова переводы приписываются к документу.

Одним из методов оценки качества извлечения опорных слов было выявление количества нежелательных для перевода слов (омонимов), которые попадали в их состав при просмотре документов различной длины, созданных с помощью двух различных алгоритмов. Для этого брались 1000 произвольно выбранных документов, которые имели разную длину. Происходило вычисление опорных слов сразу по двум алгоритмам и с помощью экспертных оценок высчитывалось общее количество омонимов. В результате было обнаружено, что классический подход TF IDF ведет себя стабильно на относительно малых документах, содержащих 100-150 слов. При превышении данного предела в опорные слова попадали нежелательные омонимы. В отличие от TF IDF, применение Okapi BM25 не привело к извлечению нежелательных слов практически на всех документах.

Поиск дубликатов в рамках рассматриваемого подхода осуществляется в два этапа. На первом этапе проводится поиск дубликатов документов, приведенных на разных языках, с помощью системы InfoStream. Системе подавалось пять опорных слов с украинских документов, представляющих собой переведенные опорные слова с украинского на русский язык (рис. 55).

Далее проводится сравнение представленных опорных слов с двенадцатью опорными словами документов, приведенных на русском языке. После этого проводилась фильтрация нежелательных, «неполных» дубликатов документов.

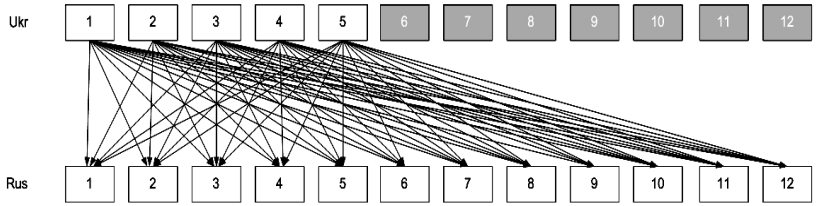


Рис. 55 – Сравнение опорных слов

Для этого были использованы следующие дополнительные критерии отсеивания не полных дубликатов: общее количество слов в переведенном варианте не должно отличаться более чем на 10 %; количество слов, которые начинаются с большой буквы (не в начале строки), не должно отличаться более чем на 3 слова; количество чисел в документах не должно отличаться более чем на два; найденные числа в документах не должны отличаться более чем на 15 %.

В результате поиска дублей новостных документов был создан параллельный двуязычный корпус документов объемом примерно 30 тыс. документов (рис. 56) [Lande, 2008].

Из полученного корпуса документов было выбрано 1000 случайных документов, подвергшихся изучению экспертами-аналитиками. Анализ показал, что в среднем 98 % содержания каждого документа имеют различные дополнения и изменения: например, ссылки на другое издательство, или другой заголовок. Также анализ показал, что из 1000 избранных документов нашелся один документ, который не совсем соответствовал документу на другом языке. Отличие состояло лишь в том, что в документе перевода были более подробно описаны подробности первичной статьи, а длина первичной статьи была небольшой – около 40 слов.

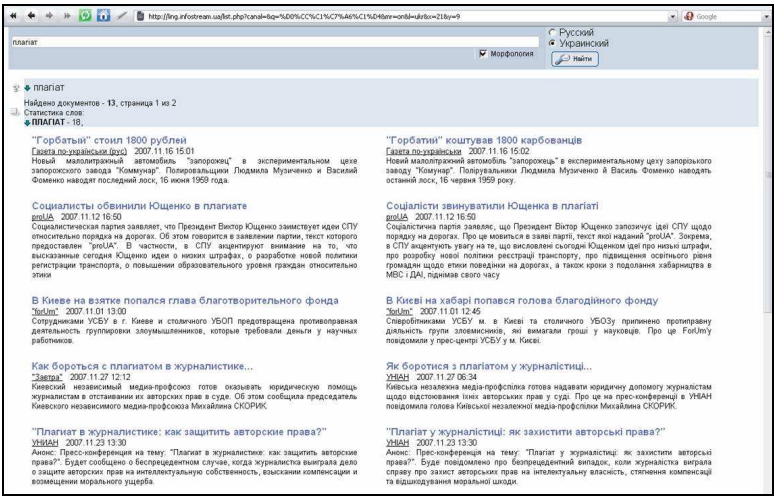


Рис. 56 – Интерфейс системы поиска в двуязычном корпусе

Приведенный алгоритм позволяет проводить поиск дубликатов, представленных не только языком, которым был написан первоначальный документ, но и на другом языке. То есть, используя механизм подключения других языков в систему, возможен поиск дубликатов, представленных сразу несколькими языками.

Как пример поиска дубликатов, авторами был создан двуязычный параллельный корпус, который в настоящее время насчитывает более 2,6 тыс. пар предложений.

На практике не всегда можно обнаружить дубликат документа, если он был создан на основе объединения нескольких текстов, эта ситуация чаще всего возникает при поиске плагиата. В таких случаях рассмотренный алгоритм, если его применять без модификаций, будет малоэффективным, но ситуацию можно улучшить путем ведения поиска опорных слов не на уровне всего текста, а на уровне нескольких абзацев. После этого этот алгоритм можно применять без ограничений.

Для оценки качества обнаружения дубликатов использовались также и формальные методы. Так на рис. 57 изображены графики, которые показывают коэффициенты близости и различия опорных слов параллельных документов. Коэффициент близости вычисляется по формуле:

$$k_1 = \frac{N_{gs}}{b}$$

где N_{gs} – количество общих ключевых слов в параллельных документах, приведенных на украинском и русском языках; b – максимальное количество одинаковых опорных слов, равное 12.

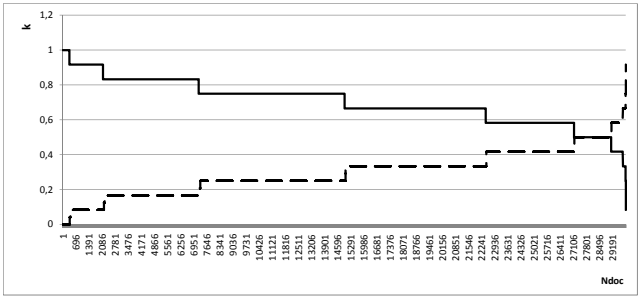


Рис. 57 – Ранжированный список коэффициентов близости (сплошная линия) и различия (пунктирная линия) документов, представленных на русском языке и их дубликатов украинском языке

Коэффициент отличия:

$$k_2 = \frac{N_{os}}{c}$$

где N_{os} – количество отличных опорных слов в документах; c – максимальное количество различных опорных слов в обоих документах, которая принимается равной 24.

Из рис. 57 видно, что пересечение графиков происходит при $k_1 = 0.5$ и $k_2 = 0.5$. При этом среднее значение общих опорных слов при поиске русскоязычных документов, близких к украиноязычным, составляет 8,45. Среднее значение общих опорных слов при поиске украиноязычных документов, близких к русскоязычным, составляет 8,97.

Примерный коэффициент ошибки вычисления опорных слов составляет 0.52. Этот коэффициент вычисляется по формуле:

$$E = \frac{\sum_1^N | R_1 - R_2 |}{N},$$

где N – общее количество документов; R_1 – количество общих опорных слов при поиске дублей русскоязычных документов, подобных украиноязычным; R_2 – количество общих опорных слов при поиске дублей украиноязычных документов, подобных русскоязычным.

В процессе поиска опорных слов для документов в результаты попадали такие опорные слова, для которых не было пары в параллельном документе или же это слово было переведено как синоним.

Анализируя параллельный корпус, было определено, что наиболее переведенными документами с русского на украинский языки и наоборот были документы, изданные теми же источниками. В частности, можно привести агентства УкрИнформ и УНИАН, «Газета по украински», «Утро Украина», которые издают новости сразу на нескольких языках.

Как показала статистика, не все параллельные документы издаются одним источником. Бывают и такие документы, которые переведены на другой язык другим издательством, чем первоисточник.

Почти все издательства, которые писали на украинском языке, имеют свое место в рейтинге

издательств, которые писали на русском языке.

Представленные алгоритмы и подходы в настоящее время используются в системе контент-мониторинга InfoStream, в частности, на этапе индексирования документа в этой системе к нему приписывается несколько значимых слов, которые переводятся на другие языки с помощью словарей переводов. Для поиска дубликатов берется несколько из найденных опорных слов из исходного документа и сравниваются со всеми переведенными опорными словами других документов.

Используя механизм подключения к системе контент-мониторинга различных языков, можно находить подобные документы или дубликаты в многоязычных базах данных, решать проблемы тематического поиска, а также поиска перепечаток.

Современные системы машинного перевода предназначены для решения ряда задач, имеющих важное значение для таких приложений, как автоматическое обнаружение опорных (ключевых) слов в документах, создание словарей, обнаружение дубликатов (плагиата), представленных на разных языках, создание массивов различных языковых версий одних и тех же документов, и, наконец, создание автоматических онлайн-переводчиков.

Сегодня практически всем известны онлайн–службы, обеспечивающие быстрый и бесплатный перевод фрагментов текстовых документов. В частности, Google (http://translate.google.com) позволяет перевод с 57 языков. Самым большим конкурентом Google является сетевая служба Bing Translator (http://www.microsofttranslator.com/), за которой стоит корпорация Microsoft, которая обеспечивает в настоящее время перевод с двадцати языков. При этом оба сетевых гиганта используют такую ветвь технологий машинного перевода, как статистический машинный перевод, которая базируется на гигантских объемах информационных ресурсов и простых и эффективных алгоритмах [Hutchins, 2005], [Hutchins, 2007].

Для систем статистического перевода характерно использование массивов текстов, представленных одновременно двумя языковыми версиями (так называемых параллельных корпусов). Чем больше объем параллельного корпуса, а так же чем качественнее перевод текстов, содержащихся в нем, тем лучше переводит статистический переводчик.

В качестве теоретической основы технологии статистического машинного перевода используется модель, базирующаяся на теореме Байеса. Данная модель предоставляет возможности улучшить перевод, используя наиболее частотные словоупотребления на разных языках, соблюдая частоты при переводе документа. Теорема Байеса в данном случае выражается простой формулой:

$$p(e \mid f) = \frac{p(e)(f \mid e)}{p(f)},$$

где f – определенный фрагмент оригинала (слово, n слов, следующих одно за другим, предложения и т.д.), e – фрагмент перевода, $p(e \mid f)$ – условная вероятность того, что переводом исходного фрагмента f будет фрагмент e, $p(f \mid e)$ – условная вероятность того, что переводу e соответствует выходной фрагмент f.

Используя формулу Байеса, формально записывается правило нахождения наиболее вероятного перевода:

$$\arg\max_e p(e \mid f) = \arg\max_e p(e)(f \mid e).$$

В приведенном выражении была проигнорирована вероятность $p(f)$, потому, что она одинакова для любого исходного текста e. В идеале вероятность $p(e)$ пропорциональна тому, насколько велика частота появления конкретного фрагмента текста в массиве, который представлен на языке перевода. Вероятность

$p(f|e)$ соответствует модели перевода. Грубо говоря, чем больше исходных текстов переводится в конкретный фрагмент *e*, тем хуже качество перевода. Вероятность $p(e)$ определяется довольно легко – по массиву возможных фрагментов перевода. По сути, такой подход позволяет разделить задачи на две части – сначала применить модель поиска подходящих предложений на языке перевода, в которых упоминаются те же понятия, что и на языке оригинала, а затем воспользоваться частотной оценкой вероятности $p(e)$, чтобы выбрать наиболее подходящий вариант перевода.

Общепринятая методология создания систем статистического машинного перевода охватывает следующие основные этапы (рис. 58):

- создание корпуса параллельных документов;
- создание корпуса параллельных предложений;
- создание массивов параллельных *N*-грамм;
- создание индексных файлов системы перевода, основанный на *N*-граммах;
- непосредственное создание модулей статистического переводчика.

В качестве источников данных для создания статистических переводчиков используют параллельные текстовые корпуса, содержащие различные языковые версии одних и тех же документов. Источниками таких документов являются сборники переведенной художественной литературы известных писателей, тексты парламентских заседаний стран и организаций, например, парламентские отчеты Канады выдаются на двух языках, официальные документы Европейского экономического сообщества выдаются 11 языках и т.д.

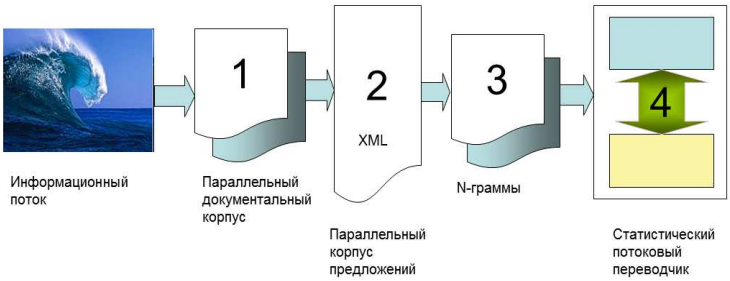


Рис. 58 – Данные, соответствующие четырем этапам формирования статистической системы машинного перевода

Для создания параллельных и мультипараллельных корпусов также используются сообщения информационных агентств, страницы веб-сайтов, имеющих несколько языковых версий.

При построении параллельных документальных корпусов для обеспечения большей точности используются дополнительные критерии, например, подсчитывается количество предложений, цифр, названий, длины фрагментов текстов и т.п.

Выравнивание документальных корпусов на уровне предложений, т.е. построение параллельных корпусов предложений, выполняется на основе главного постулата систем статистического перевода – принципа монотонности [Потемкин, 2008]. Этот принцип заключается в том, что различные языковые версии одного и того же документа содержат предложения, размещенные в одном и том же порядке, то есть второе предложение следует после первого, третье – после второго и т.д.

Следующим этапом формирования базы данных статистического переводчика является формирование массива N-грамм. N-граммой называется последовательность из N слов одного текста, идущих друг за другом. Например, во фразе «простейшей моделью перевода является дословный перевод»

содержатся такие триграммы:

- «простейшей моделью перевода»;
- «моделью перевода есть»;
- «перевода является дословный»;
- «есть дословный перевод».

N-грамма – это традиционный объект исследования компьютерных лингвистов. Первое практическое применение N-граммы получили в программах определения языков, которыми написаны тексты, проверки правописания, а затем уже в технологиях статистического машинного перевода.

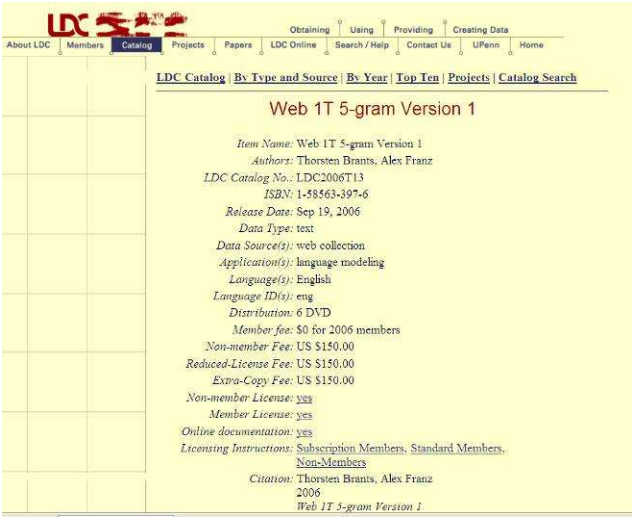


Рис. 59 – Фрагмент описания корпуса пентаграмм компании LDC

N-граммы, соответствующие одноязычным текстовым корпусам, сегодня являются коммерческими продуктами, создаются и предлагаются на рынке. Например, массив английских 5 грамм (пентаграмм) предлагается компанией LDC (рис. 59). Этот массив содержит 5 триллионов записей и размещается в архивированном виде на 6 DVD-дисках. Технологии N-

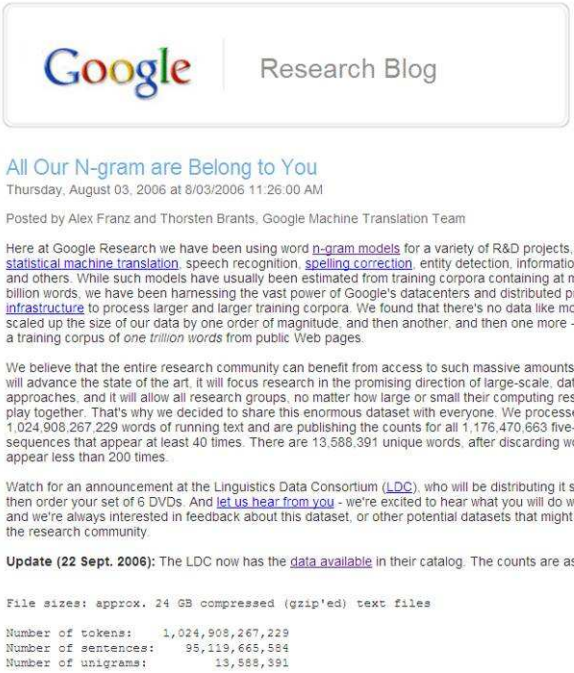
грамм, кроме того, широко используются и пропагандируются компаниями Google (рис. 60) и Bing (рис. 61).

При построении баз данных современных систем перевода создаются массивы N-грамм (чаще – пентаграмм).

Для этих массивов в рамках технологий статистического машинного перевода используются параллельные двуязычные корпуса предложений. Для каждой пары предложений строятся N-граммы на одном языке, которым соответствуют (по месту в соответствующем предложении) N-граммы на другом языке. Например, если представить предложения, в котором функцию слов выполняют латинские буквы: «abcdefgh», а переводом (соответствующим предложением из параллельного корпуса) будет: «а б в г д е ж з», то для них будут построены пары пентаграмм:

$$П_{1eng}(abcde)\sim П_{1rus}(абвгд),\ П_{2eng}(bcdef)\sim П_{2rus}(бвгде),$$
$$П_{3eng}(cdefg)\sim П_{3rus}(вгдеж),\ \ldots$$

Далее проводился подсчет количества N-грамм, которые встречаются в параллельном корпусе предложений. В случае, если на одном языке N-грамме соответствует несколько N-грамм на другой язык, то выбирается наиболее частотная N-грамма.



Google Research Blog

All Our N-gram are Belong to You

Thursday, August 03, 2006 at 8/03/2006 11:26:00 AM

Posted by Alex Franz and Thorsten Brants, Google Machine Translation Team

Here at Google Research we have been using word n-gram models for a variety of R&D projects, such as statistical machine translation, speech recognition, spelling correction, entity detection, information extraction, and others. While such models have usually been estimated from training corpora containing at most a few billion words, we have been harnessing the vast power of Google's datacenters and distributed processing infrastructure to process larger and larger training corpora. We found that there's no data like more data, and scaled up the size of our data by one order of magnitude, and then another, and then one more - resulting in a training corpus of *one trillion words* from public Web pages.

We believe that the entire research community can benefit from access to such massive amounts of data. It will advance the state of the art, it will focus research in the promising direction of large-scale, data-driven approaches, and it will allow all research groups, no matter how large or small their computing resources, to play together. That's why we decided to share this enormous dataset with everyone. We processed 1,024,908,267,229 words of running text and are publishing the counts for all 1,176,470,663 five-word sequences that appear at least 40 times. There are 13,588,391 unique words, after discarding words that appear less than 200 times.

Watch for an announcement at the Linguistics Data Consortium (LDC), who will be distributing it soon, and then order your set of 6 DVDs. And let us hear from you - we're excited to hear what you will do with the data, and we're always interested in feedback about this dataset, or other potential datasets that might be useful for the research community.

Update (22 Sept. 2006): The LDC now has the data available in their catalog. The counts are as follows:

```
File sizes: approx. 24 GB compressed (gzip'ed) text files

Number of tokens:    1,024,908,267,229
Number of sentences:    95,119,665,584
Number of unigrams:         13,588,391
```

Рис. 60 – Корпус N-грамм в технологиях Google

Типичный алгоритм работы статистического переводчика следующий. На вход модуля перевода подается документ на языке оригинала, который сразу же разбивается на предложения. Для каждого предложения строятся N -граммы и выполняется поиск их переводов на другой язык. В случае если перевода какой-то N-граммы не удается обнаружить, ищется соответствующая (N-1)-грамма и выполняется поиск ее перевода и т.д. Если приложение не выявляло перевода для биграмм, выполнялся поиск слова в словаре переводов слов. Если возникает ситуация, что некоторое слово переводится и таким образом, то оно остается на языке оригинала или (в некоторых системах) проводился «псевдоперевод» (транслитерация и т.п.).

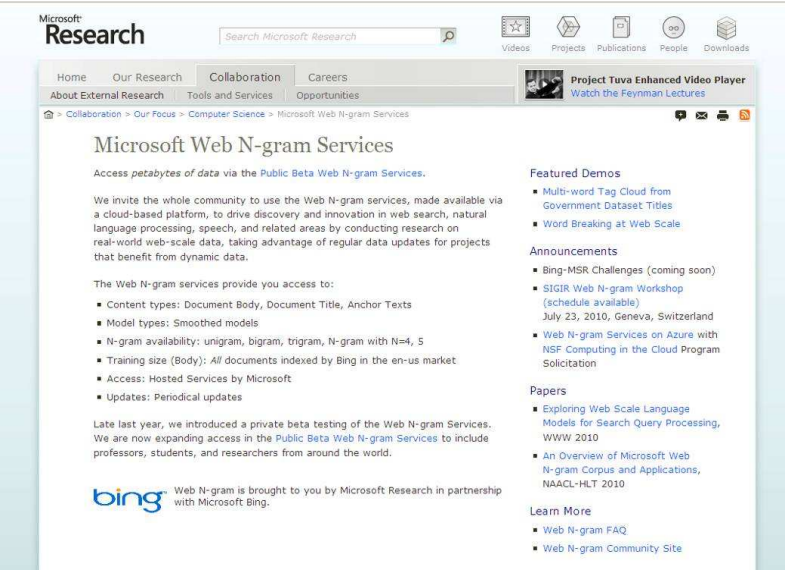


Рис. 61 – N-граммы в технологии Bing Translate

После перевода всех предложений документ форматировался по шаблону и выводился пользователю как результат.

На сегодняшний день реализовано и доступно пользователям несколько статистических переводчиков, в основе которых заложена технология перевода с помощью N-грамм. Это уже названные ранее Google Translate, Bing Translator и еще не функционирующий, но уже разрекламированный статистический переводчик компании IBM, часть функциональных возможностей которого внедрена на веб-сайте компании в рамках технологии n.Fluent с целью упрощения общения разработчиков разных языках.

При построении первичного параллельного документального корпуса в подсистеме машинного перевода системы InfoStream, в отличие от общеизвестных систем, использовались лингвостатистические алгоритмы, применяемые к результатам контент-мониторинга сетевых СМИ –

массивов новостей.

В основу алгоритма определения параллельных документов был взят метод сравнения опорных слов. На этом этапе были построены статистические морфологические словари с использованием документов из системы контент-мониторинга InfoStream.

Для дальнейшей работы со словарями были взяты только существительные. Имеющиеся словари были дополнены именами и названиями компаний и организаций. Также были созданы словари переводов, представлявшие собой частотные словари переводов существительных.

Опорные слова вычислялись по формуле Okapi BM25 с использованием морфологического статистического словаря. Для определения подобных документов, приведенных на разных языках, были взяты выделенные опорные слова и переводы этих слов на другой язык с помощью словаря переводов. Системе подавалось пять переведенных ключевых слов с украинского на русский язык с украинских документов. Далее проводилось сравнение представленных ключевых слов с двенадцатью ключевыми словами документов на русском языке. В случае если документы по заданному критерию были найдены, данные документы считались подобными. Далее проводился поиск дублей документов. Были использованы следующие дополнительные критерии отсеивания неполных дубликатов, представленных на разных языках:

- общее количество слов в оригинале и переводе не должно отличаться более, чем на 10 % ;
- количество слов, начинающихся с большой буквы, не должно отличаться более чем на 3 ;
- количество чисел в документах не должно отличаться более чем на два и т.д.

При построении конечного корпуса параллельных документов из первичного набора параллельных

документов проводилась фильтрация и отсеивание лишних дублей. В итоге, выбирался тот русский документ, который наиболее соответствует количеству слов в украинском документе. На основании приведенного алгоритма был создан параллельный украино-русский корпус документов.

Следующий шаг к созданию параллельного переводчика – выравнивание параллельного русско-украинского корпуса документов на уровень предложений. При этом предполагалось, что предложения заканчиваются символами «.», «!», «?», а также с учетом того, что сокращения или инициалы с точками не обязательно определяют конец предложения.

Посл этого проводился подсчет количества предложений в параллельных документах. Если данные документы по количеству предложений, были одинаковыми, то эти предложения использовались в дальнейшей обработке. Каждое предложение было разделено на слова. Под словом подразумевалось любое сочетание символов, отделенное от других групп символов пробелом. Также накладывались дополнительные ограничения на определение слова в каждом из языков. Например, в словах на украинском языке, перед которыми упоминались слова: который, которая, что который т.д., определялись как сложные и рассматривались как одно слово. Далее проводился подсчет количества слов в параллельных предложениях. В результате, в параллельный корпус предложений отбирались только те предложения, которые по количеству слов не отличались более чем на одно слово.

Следующий этап построения статистического переводчика заключался в создании статистических словарей триграмм (рис. 62), биграмм и статистического словаря переводов слов из параллельного корпуса предложений, который был построен раньше. Для каждой пары предложений были построены триграммы на одном языке, которые отвечали триграммами на другом языке.

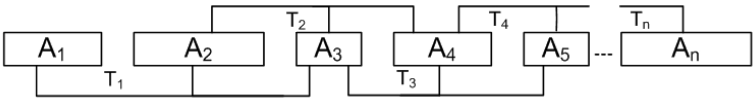


Рис. 62 – Разделение предложения на триграммы
(A_1–A_n) – слова, (T_1–T_n) – триграммы

Для определения слова, входящего в состав N-грамм, использовались те же правила разделения предложения на слова, которые применялись для выравнивания параллельного корпуса. В основу словарей вошли пары N-грамм – N-грамма на одном языке и перевод ее на другой язык. Проводился подсчет количества N-грамм, которые встречались в параллельном корпусе предложений. В случае, если на одном языке N-грамме соответствовало несколько N-грамм на другом языке, то выбирается наиболее частотная N-грамма.

Начало технологической цепочки статистического переводчика InfoStream заключалось в том, что представленный документ делился на предложения. Для каждого предложения строились триграммы и выполнялся поиск перевода триграммы на параллельный язык. В том случае если перевода не удалось обнаружить, то использовалась биграма, и далее выполнялся поиск перевода данной биграммы.

Если переводчик не проявлял перевода для триграммы или биграммы, то выполнялся поиск слова в словаре переводов слов. Если происходила ситуация, когда какое-то слово известно переводчику и не складывалось ни в одном документе – в данном случае проводился псевдоперевод слова. После перевода всех предложений документ форматировался по шаблону как входящий документ и выводился пользователю как результат.

Основные проблемы, с которыми авторы столкнулись при построении словарей, были большие объемы словарей и обеспечение быстрого поиска по этим массивам данных. На данный момент общий

объем словарей занимает более 2 Гб, в таком случае прямой поиск в данном словаре не эффективен. Поэтому для поиска в словарях триграмм и биграм был применен метод бинарного поиска.

Для осуществления бинарного поиска был построен индексный файл для каждого словаря. В качесте ключа был взят хэш от каждого слова, а в качестве функции хеширования – функция CRC-32, которая возвращает 4-х байтовое положительное число. При этом подходе очевидны коллизии, представляющие собой одинаковые хэш-значения различных по написанию слов. Для преодоления коллизий использовался специальный словарь для разрешения коллизий, который в свою очередь загружался в оперативную память.

В отличие от пятиграмм, используемых в Google и Bing Translator, для построения российско-украинского переводчика оказалось достаточным использовать триграммы, что значительно повысило производительность системы. В результате сравнительного перевода одного и того же фрагмента текста, взятого из системы контент-мониторинга InfoStream, был сделан вывод, что на сегодняшний день статистические переводчики немного проигрывают в качестве перевода переводчикам, основанным на правилах.

Предложенная методология позволила создать систему с элементами самообучения (после работ, сделанных на этапе инициализации) статистического перевода, ориентированную на массовый перевод текстовой информации из информационных потоков, представленных на российском и украинском языках (рис. 63).

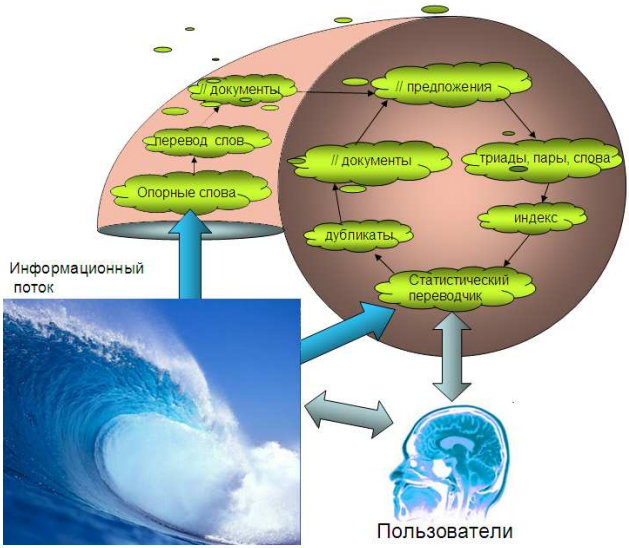


Рис. 63 – Статистический переводчик – система с обратной связью

Документы из информационного потока поступают в модуль перевода, осуществляется выявление дубликатов, представленных на разных языках, дополняется параллельный документальный корпус, затем происходит выравнивание на уровне предложений, построение массивов триграмм, биграмм и слов, формируется индекс обновленного переводчика, который используется модулем перевода, на вход которого поступают как запросы отдельных пользователей (в режиме онлайн), так и весь сканированный системой InfoStream информационный поток.

Сегодня можно констатировать, что проблема создания приемлемых по качеству (обеспечивающих понимание текстов пользователями) онлайн-переводчиков решена. Кроме того, становится очевидным, что именно ветвь из систем статистического машинного перевода является наиболее эффективной и качественной.

Вместе с тем, всем системам статистического перевода в той или иной степени присущ ряд проблем, среди которых можно назвать:

1. Нехватка необходимого объема исходных переводов (параллельных текстовых корпусов).

2. Практически все рафинированные системы статистического машинного перевода некачественно обрабатывают длинные фразы, превышающие 10-12 слов.

3. Проблемы скорости, которые не позволяют создавать полноценные поточные переводчики.

4. Можно признать, что для обеспечения качества перевода необходимы хотя бы элементарные правила. Разработчики систем машинного перевода для улучшения качества перевода вводят некоторые «сквозные» правила, тем самым превращая чисто статистические системы в гибридные.

Безусловно, добавление некоторых правил, то есть создание гибридных систем, несколько улучшает качество переводов, особенно при недостаточном объеме входных данных, используемых при построении индекса машинного переводчика.

Перед системами машинного перевода возникает ряд актуальных задач, диктуемых насущными потребностями пользователей:

- Повышение точности и адекватности перевода. Как показано многочисленными примерами, решение задачи достигается путем увеличения объемов соответствующих баз данных параллельных текстов, а также учета достаточно небольшого набора правил для каждой из пар языков.

- Перевод полных документов, а не только фрагментов. Теоретически современные системы статистического машинного перевода способны обеспечить этот сервис,

однако для предоставления его в массовом масштабе необходимо достичь высокой производительности программ-переводчиков, не всегда возможно. Вторая причина – коммерческая – полные версии переводчиков предоставляются зарегистрированным пользователям за плату.

- Перевод документальных потоков, а не отдельных документов. Подобный сервис востребован, например, при создании языковых версий сайтов.

- Нахождение информационных дубликатов, представленных на разных языках. Решение этой проблемы обеспечит, например, возможность анализа широты охвата PR-кампании, границы проведения информационных операций, выявление перепечаток, плагиата.

- Создание информационно-поисковых систем с интерфейсом на языке пользователя, которые предоставляют результаты на этом же языке, но обеспечивают поиск в иноязычных сегментах веб-пространства. Например, создание украиноязычной поисковой системы по китайским веб-ресурсам (сегодня из-за языковых проблем огромный китайский сегмент веб-пространства является так называемым «скрытым веб» (Deep Web) для нашего пользователя).

- Предоставление возможности автоматической подготовки аналитических документов на основе информации, представленной на разных языках. Эта проблема стыкуется с проблемой автоматического реферирования, обобщения документов, однако относится к обработке исходной информации, представленной на разных языках.

- Генерация систем, обеспечивающих правильное звуковое воспроизведение переведенной информации. Такие системы могут быть полезными пользователям с ограниченными возможностями, а также пользователям, которые нуждаются в информации в экстремальных условиях (например, водителям автотранспорта, путешествующим за границей, сотрудникам правоохранительных органов и т.п.).

Для частичного решения задач обнаружения спам-сообщений и новых событий реализован отдельный режим, близкий по идеологии к режиму «поиска подобных документов» в системе контент-мониторинга InfoStream. В рамках системы InfoStream сообщение считается подобным исходному, если оно содержит определенное количество (a) наиболее значимых термов (опорных слов) из него (этот критерий назван авторами a-подобием).

Под спам-популярностью сообщения будем понимать количество a-подобных ему сообщений в текстовом корпусе спама. Под СМИ-популярностью понимается количество a-подобных сообщений в ретроспективной базе электронных СМИ. Массив сообщений, заведомо точно определенных авторами как спам, был ранжирован по спам-популярности. Полученная зависимость «спам-популярность – количество сообщений» оказалась близкой к гиперболической. Для каждого из сообщений, ранжированных по значению СМИ-популярности, также построена зависимость «СМИ-популярность – количество сообщений». Было обнаружено некоторое количество сообщений, которые характеризуются большим соотношением спам-популярности к СМИ-популярности. Этот факт позволяет судить о совокупности терминов, определяющих спам-популярность, как о еще одном фильтре, который можно реализовать в антиспамовском программном обеспечении. Сообщения, в которых СМИ-популярность превышает спам-популярность, но все же

являющиеся спамом, оказались несанкционированными рассылками информационно-аналитических материалов, представляющих определенный интерес для информационных агентств.

В представленном подходе к выявлению спама существенно то, что определяется близость исследуемого сообщения не только корпусу спама, но и корпусу электронных СМИ.

Итак, под локальной популярностью сообщения в информационном потоке понимается количество *a*-подобных сообщений ему в тот период (час, день, неделю), когда появилось исходное сообщение. Под глобальной популярностью понимается количество *a*-подобных сообщений за значительный ретроспективный период.

Особый интерес представляют сообщения, характеризующиеся большим соотношением локальной популярности к глобальной. Этот факт позволяет судить о событиях, описываемых в данных сообщениях, как о новых. Таким образом получен алгоритм обнаружения документов, получивших популярность только в последнее время, что является отдельным решением актуальной проблемы выявления новых событий [Снарский, 2007].

Общепринятая технологическая схема решения задачи обнаружения новых событий из потока новостей, как правило, предполагает, что новые события описываются в документах, для которых с помощью отдельных программных модулей во временной ретроспективе формируются цепочки подобных документов (сюжетные цепочки). Документы, отражающие различные новые события, могут быть основой новых групп взаимосвязанных документов-кластеров (группировка событий). В свою очередь, каждый из этих кластеров со временем может стать основой формирования полноценной сюжетной цепочки.

Приведем некоторые предположения, относящиеся

к документам, которые содержат информацию о новых событиях:

а) минимальное время, прошедшее с момента публикации документа;

б) близость лексического состава документа к лексическому составу массива документов за небольшой промежуток времени (оперативность новостей);

в) существенное различие лексического состава документа от лексического состава массива документов за значительный период времени – окна наблюдения;

г) наличие в документе терминов, входящих в плюс-словарь, который включает важные для содержания новостей слова типа «теракт», «конфликт», «сенсация» и т.п.);

д) высокий ранг репутации источника, а также допустимость лексики заголовков новостей (определяемого экспертами);

е) отсутствие дублирования информации.

Введем обозначения: N – величина окна наблюдения потока новостей; n – величина массива оперативных новостей (($n < N$); D_i – i-й документ; PlusDic – плюс-словарь; $sim(D_i, D_j)$ – мера близости документа i документа j; sim (Di, PlusDic) – мера близости документа i плюс-словаре; $Rangi$ – ранг источника, который соответствует i-му документу.

В этих обозначениях мера близости лексического состава документа и лексического состава массива оперативных новостей рассчитывается следующим образом:

$$\sum_{j=1}^{n} sim(D_i, D_j),$$

Соответственно мера близости лексического состава документа от лексического состава массива

других документов из окна наблюдения рассчитывается следующим образом:

$$\sum_{j=n}^{N} sim(D_i, D_j).$$

При этом степень близости отдельных документов, рассчитывается как введенное выше α-подобие.

Формула для расчета ранга документа как «носителя» информации о новых событий с учетом условий а)-е) может быть записана следующим образом:

$$N(D_i) = \frac{Rang_i \cdot sim(D_i, PlusDic) \cdot \sum_{j=1}^{n} sim(D_i, D_j)}{\log(i+1) \cdot \sum_{j=n}^{N} sim(D_i, D_j)},$$

с учетом приведенных выше обозначений, а также того, что если нумерация документов из потока производится в обратном порядке, значение $\log(i+1)$ в знаменателе отражает вклад времени, прошедшего с момента публикации события.

На основе приведенной формулы может происходить ранжирование документов, поступающих в системы интеграции новостей. Построенный в рамках представленной технологии алгоритм используется в настоящее время в системе контент-мониторинга InfoStream, на вход которой ежедневно поступает более 80 тыс. документов.

Данный алгоритм реализует прогнозно-аналитическую модель, основная методология оценки достоверности которой в настоящее время заключается в экспертном сравнении выявленных новых событий с основными сюжетами, полученными через определенный интервал времени. Для настройки алгоритма экспертами использовались такие «рычаги», как параметры *N, n*, плюс-словарь, массив рангов источников информации, массив исключений для заголовков и адресов.

В настоящее время во многих популярных системах интеграции новостей задача выявления новых событий заменяется выявлением основных новостных сюжетных цепочек. Такой подход, конечно, частично решает названную задачу, однако, предоставляя пользователям ответ на вопрос «о чем больше пишут в последнее время», фактически отличается целевой функцией.

Было проведено ретроспективное исследование для оценки, насколько сегодняшние события, определяемые в соответствии с предложенным подходом, станут основой сюжетов на следующий день. Оказалось, что таких событий не более 20%. Зачастую большая часть сюжетов на следующий день повторяет сюжеты дня предыдущего. Приходится признавать, что не все новые события одинаковы по важности и порождают в дальнейшем значительные кластеры подобных документов.

Предложенный подход нельзя считать окончательным решением поставленной задачи. Например, не всегда изменение размеров окон наблюдения и объемов оперативных массивов может привести к адекватному выявлениюю новых обстоятельств, которые имеют свою предысторию. Рассматриваемый плюс-словарь требует постоянного сопровождения, а в некоторых случаях «персонализации».

Однако, полученные практические результаты показали свою эффективность как существенное дополнение к поисковым режимам. При этом самое важное, пожалуй то, что пользователь привязывается не к новым сообщениям, а к новым событиям реального мира.

Задача выявления нечетких дубликатов является одной из актуальных и сложных, особое практическое значение она принимает, в частности, при интеграции информационных ресурсов, решении задачи борьбы с плагиатом, определения спам-рассылок и т.п. Серьезное упрощение названной задачи может быть получено за

счет применения формальных методов, например, математической статистики, сигнатурных алгоритмов, репутационных подходов (например, путем ранжирования первоисточников, тематик, опорных слов и т.д.).

На формальном уровне путем сопоставления фрагментов текстов, шинглов, лингвистических сигнатур и т.п. сегодня с успехом используются нечеткие дубликаты, которые формируются путем копирования, прямого перевода с иностранных языков, компиляции и т.п.

Однако нечеткие дубликаты включают также результаты переработки оригиналов на содержательном уровне, например, переводы одних и тех же событий, текстов, идей, описание различных аспектов разными авторами. Кроме того, нечеткие дубликаты могут быть представлены в различных медиа-средах. В данный момент сходство обобщенных таким образом документов не всегда может быть определена с помощью приведенных выше методов и алгоритмов. Именно поэтому сравнительный анализ электронных текстов представляет собой ныне открытую научно-практическую проблему.

3.9.3. Упорядочивание информации

Основным критерием упорядочения информации в современных ИПС – ранжирование. Ранжирование текстовых и гипертекстовых документов существенно различается. Текстовые документы могут ранжироваться по уровням релевантности и другим параметрам, экстрагируемым из текстов. Ранжирование гипертекстовых документов возможно также по свойствам, обусловливается сетевой структурой, гиперссылками.

Например, для определения авторитетности веб-страницы как источника информации или посредника, используется анализ графа, созданного веб-документами и соответствующими гиперссылками. Два самых известных алгоритмов ранжирования веб-

страниц, основанных на связях HITS (hyperlink induced topic search) и PageRank, были разработаны в 1996 году в IBM Дж. Клейнбергом (JM Kleinberg) и в Стэнфордском Университете С. Брином (S. Brin) и Л. Пейджем (L. Page).

Оба алгоритма предназначены для решения «проблемы избыточности», свойственной широким запросам, увеличение точности результатов поиска на основе методов анализа сложных сетей.

Алгоритм HITS

Алгоритм HITS (Hyperlink Induced Topic Search), предложенный Дж. Клейнбергом, является реализацией латентно-семантического индексирования для ранжирования выдачи информационно-поисковых систем. Алгоритм HITS обеспечивает выбор из информационного массива лучших «авторов» (первоисточников, на которые осуществляют ссылки) и «посредников» (документов, от которых идут ссылки цитирования). Страница является хорошим посредником, если она содержит ссылки на ценные первоисточники, и наоборот, страница является хорошим автором, если она упоминается хорошими посредниками.

Для каждого документа $d_j \in D$ рекурсивно вычисляется его значимость как автора $a(d_j)$ и посредника $h(d_j)$ по формулам:

$$a(d_j) = \sum_{i=1, i \neq j}^{|D|} h(d_i), \;\; h(d_j) = \sum_{i=1, i \neq j}^{|D|} a(d_i).$$

Покажем, что алгоритм HITS подбен алгоритму кластеризации LSA. Введем понятие матрицы инциденций A, элемент которой $a_{i,j}$ равен единице, когда документ d_i содержит ссылку на документ d_j, и нулю в противном случае. Воспользуемся сингулярным расписанием: $A = USV^T$, где A – квадратная

диагональная матрица с неотрицательными диагональными элементами $s_{i,j}$. Рассмотрим матрицу $A^T A$, для которой справедливо: $A^T A = VSU^T USV^T = VS^2V^T$, где S^2 – диагональная матрица с элементами $s_{i,i}^2$. Соответственно, для матрицы AA^T будет справедливо $AA^T = US^2U^T$. Очевидно, что как и случае с алгоритмом LSA, собственные векторы, соответствующие самым сингулярным значениям AA^T (или $A^T A$) будут соответствовать статистически наиболее важным авторам (или посредникам).

Алгоритм вычисления рангов HITS приводит к росту рангов документов при увеличении количества и степени связанности документов соответствующего сообщества. В этом случае в результаты поиска системы, использующей алгоритм HITS, могут попасть в большом количестве документы по темам, отличным от информационной потребности пользователя, но тесно связанных между собой, то есть часть выдаваемых результатов может отклониться от доминирующей тематики. В этом случае происходит так называемый сдвиг тематики (topic drift) за счет наличия «тесно связанных сообществ» документов (Tightly-Knit Community, TKC).

Для решения этой проблемы как некоторое расширение стандартного алгоритма HITS был предложен вероятностный алгоритм PHITS (Probabilistic HITS). В рамках этого метода предполагается: D – множество цитирующих документов, C – множество ссылок, Z – множество классов (факторов). Предполагается также, что событие $d \in D$ происходит с вероятностью $P(d)$.

Условные вероятности $P(c \mid z)$ и $P(z \mid d)$ используются для описания зависимостей между

наличием ссылки $c \in C$, латентным фактором $z \in Z$ и документом $d \in D$.

Оценивается функция правдоподобия:

$$L(D,C) = \prod_{c \in C, d \in D} P(d,c) = \prod_{c \in C, d \in D} P(d)P(c \mid d),$$

где

$$P(c \mid d) = \sum_{z \in Z} P(c \mid z)P(z \mid d).$$

Цель вероятностного алгоритма PHITS заключается в том, чтобы подобрать такие $P(z)$, $P(c \mid z)$, $P(d \mid z)$, чтобы максимизировать $L(D,C)$.

После этого:

$P(c \mid z)$ – ранги авторов;

$P(d \mid z)$ – ранги посредников.

Для вычисления рангов необходимо задать количество факторов во множестве Z, и тогда $P(c \mid z)$ будет характеризовать качество страницы как автора в контексте тематики Z. К недостаткам метода следует отнести то, что итеративный процесс чаще останавливается не в абсолютном, а на локальном максимуме функции правдоподобия L.

Вместе с этим в ситуациях, когда в множестве найденных веб-страниц нет явного доминирования тематики запроса PHITS превосходит алгоритм HITS.

Алгоритм PageRank

Алгоритм PageRank близкий по идеологии к литературному индексу цитирования и рассчитывается для любого документа с учетом количества ссылок с других документов на данный документ. При этом PageRank по HITS, в отличие от литературного индекса цитирования, не считает все ссылки равнозначными.

Принцип расчета ранга веб-страниц PageRank

основывается на модели «случайного блуждания» пользователя по следующему алгоритму: пользователь сети Интернет открывает случайную веб-страницу, из которой переходит по случайно выбранной гиперссылке. Затем он перемещается на другую веб-страницу и снова активизирует случайную гиперссылку и т.д., постоянно переходя от страницы к странице, никогда не возвращаясь. Иногда ему такое блуждание надоедает, и он снова переходит на случайную веб-страницу – не по ссылке, а набрав вручную некоторое URL. В этом случае вероятность того, что блуждающий в WWW пользователь перейдет на некоторую определенную веб-страницу – это ее ранг. Очевидно, PageRank веб-страницы тем выше, чем больше других страниц ссылается на нее, и чем эти страницы популярнее.

Пусть имеется n страниц $\{d_1, ..., d_n\}$, которые ссылаются на данный документ (веб – страницу A), а $C(A)$ – общее число ссылок с веб-страницы A на другие документы. Определяется некоторое фиксированное значение δ как вероятность того, что пользователь, просматривая какую-либо веб-страницу из множества D, перейдет на страницу A по ссылке, а не набирая ее URL в явном виде. В рамках модели вероятность продолжения этим пользователем веб-серфинга по сети N с веб-страниц без использования гиперссылок, путем ручного ввода адреса (URL) случайной страницы составит $1-\delta$ (альтернатива перехода по гиперссылкам). Индекс PageRank $PR(A)$ для страницы A рассматривается как вероятность того, что пользователь окажется в некоторый случайный момент времени на этой странице:

$$PR(A) = (1-\delta)/N + \delta \sum_{i=1}^{n} \frac{PR(d_i)}{C(d_i)}.$$

По этой формуле индекс страницы легко подсчитывается простым итерационным алгоритмом. На практике применяется 30 шагов итерации для

достижения устойчивых результатов.

Несмотря на различия алгоритмов HITS и PageRank, в этих алгоритмах общее то, что авторитетность (вес) узла зависит от веса других узлов, а уровень «посредника» зависит от того, насколько авторитетны узлы, на которые он ссылается.

Расчет авторитетности отдельных документов сегодня широко используется в таких приложениях, как определение порядка сканирования документов в сети роботом ИПС, ранжирование результатов поиска, формирование тематических обзоров и т.п.

В настоящее время получили широкое распространение технологии искусственного повышения рангов отдельных веб-документов или их групп (сайтов) путем установления гиперссылок, не имеющих отношения к их содержанию. Эти технологии, названные методами поисковой оптимизации (SEO, Search Engine Optimization), основываются на приспособлении к существующим алгоритмам ранжирования веб-документов наиболее популярными поисковыми системами. В свою очередь, такие технологии приводят к необходимости постоянного совершенствования алгоритмов ранжирования в поисковых системах, ориентации на содержательную составляющую веб-документов при определении их рангов.

Алгоритм Salsa

Алгоритм ранжирования Salsa (стохастический подход для Link-Structure Analysis – Стохастический алгоритм анализа структуры Связей) был предложен Ш. Мораном (Ш. Moran) и Р. Лемпела (P. Lempel) как некий симбиоз алгоритмов PageRank и HITS, что позволяет сократить последствия образования ТСС – «тесно связанных сообществ» документов.

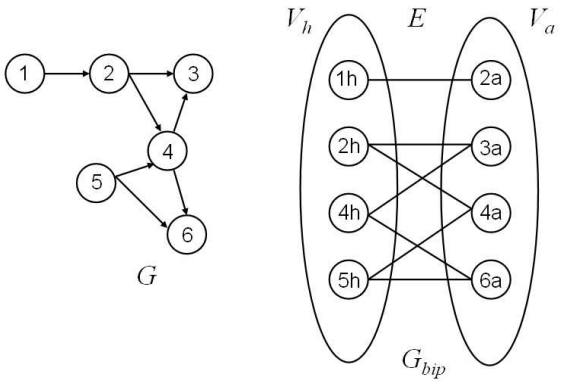
Как и в методе PageRank, в случае алгоритма Salsa предполагается модель случайного блуждания пользователя по веб-графу, однако предполагается

наличие двустороннего «серфинга». Согласно алгоритму Salsa:

1. Из произвольного узла v пользователь случайным образом возвращается к узлу u, который ссылается на узел v. Выбор узла v делается случайно, при условии, что узлы v и u принадлежат веб-графу.

2. С узла u пользователь наугад переходит к узлу w, если существует связь (u, w).

Веб-граф G (рис. 64 а) может быть преобразован в двудольный ненаправленный граф G_{bip}, (рис. 64 б) и определен как совокупность $G_{bip} = (V_h, V_a, E)$, где h обозначает посредников, V_h – совокупность узлов-посредников (тех, из которых выходят ссылки), и узлов-авторов (тех, на которые ведут ссылки). Необходимо отметить, что одни и те же узлы могут быть одновременно и авторами и посредниками.



а) *б)*

Рис. 64 – Salsa: конструкция двудольных графов

Каждая неизолированная страница $s \in G$ представлена в G_{bip} одним или двумя узлами s_h и s_a. В этом двудольном графе алгоритм ранжирования Salsa

реализует два различных случайных перехода. При каждом переходе возможно «посещение» узлов только с одной из двух частей графа G_{bip}.

Каждый путь длиной два в G_{bip} представляет собой переход по гиперсвязи (при прохождении от доли посредников к части авторов в G_{bip}), и отход вдоль гиперсвязи (при прохождении в обратном направлении). Это движение в обратном направлении напоминает танец сальса, который ассоциируется с названием данного алгоритма.

Так как посредники и авторы, относящиеся к теме t, должны быть явно выражены в G_{bip} (доступны из многих узлов благодаря прямым ссылкам или коротким путям), предполагается что авторы V_a и посредники с V_h, относящиеся к теме t, будут наиболее часто посещаемыми при случайных «блужданиях» пользователей.

В алгоритме ранжирования Salsa исследуются две различные цепи Маркова, которые ассоциируются с этими случайными блужданиями: цепь Маркова на стороне авторов G_{bip} (цепь авторов), и цепь Маркова на стороне посредников G_{bip}.

Такой подход позволяет ввести две стохастические матрицы перехода цепей Маркова, которые определяются следующим образом: строится матрица инциденций W ориентированного графа G. Обозначим через W_r – матрицу, полученную делением каждого ненулевого элемента W на сумму значений соответствующей строки, а через W_c – матрицу, полученную делением каждого ненулевого элемента W на сумму элементов в соответствующем столбце. Тогда матрица $H = W_r W_c^T$, соответствующая посредникам, будет состоять из ненулевых строк и столбцов, а

матрица авторов, сооответственно, будет состоять из ненулевых строк и столбцов $A = W_c^T W_r$.

В рамках алгоритма ранжирования Salsa игнорируются строки и столбцы матриц A и H, состоящие полностью из нулей, так как по определению, все узлы G_{bip} имеют не менее одной связи. В результате матрицы A и H используются для вычисления рангов тем же путем, что и в алгоритме HITS.

Показано [Lempel, 2000], что вероятность перехода, которая сходится в процессе итерационного процесса к узлу как к автору, имеет очень простую форму:

$$\pi_v = c_1 \cdot InDegree(v),$$

а вероятность возврата к узлу u как к посреднику:

$$\pi_u = c_2 \cdot OutDegree(u),$$

где c_1 и c_2 – некоторые константы, $InDegree$ и $OutDegree$ – это количество исходящих и входящих ссылок, соответственно.

Г. Лемпель и С. Моран продемонстрировали, что алгоритм ранжирования Salsa менее чувствителен к эффекту тесно связанных сообществ, чем алгоритм HITS, но при условии, что предварительно вручную в документах удаляются ссылки, которые не относятся к теме, которая исследуется. Это требование на практике ведет к большим затратам, в результате чего авторам пока не известно случаев использования этого алгоритма ранжирования в реально работающих системах.

4. АДАПТИВНОЕ АГРЕГИРОВАНИЕ ИНФОРМАЦИИ

Задачи мониторинга информационных потоков большого объема в компьютерных сетях усложняются отсутствием типовых методик и решений, неполнотой существующих технологических подходов.

В настоящее время исследования по проблемам мониторинга и анализа информационных потоков большого объема в компьютерных сетях носят узко специализированный характер. В то же время опыт разработки и внедрения корпоративных информационных систем свидетельствует о необходимости создания и внедрения документальных информационных хранилищ для обеспечения научных исследований, получения различных аналитических сведений, навигации в документальных информационных потоках сверхбольших объемов.

Представляется очень важным, чтобы агрегация документальной информации, формирование информационного хранилища были адаптивными, т.е. ориентированными на информационные потребности пользователей. Если учитывать динамику и объемы доступной в Интернете информации (на сегодня доступно более триллиона документов), то становится очевидным, что обеспечение эффективного доступа в режиме поиска к информации в отрыве от информационных потребностей пользователей является практически неразрешимой задачей.

Основная идея адаптивной агрегации информации заключается в сборе и сохранении в информационном хранилище только той информации, которая соответствует информационным потребностям пользователей (существующих или потенциальных) [Додонов, 2013-1]. Для этого предполагается, что по мере развития системы в ее информационное хранилище будут попадать актуальные документы из сети Интернет соответствующие текущим запросам пользователей. Естественно, с ростом количества

пользователей, объемы информационного хранилища (репозитория) будут также расти, что в конкретный момент приведет к пересмотру его состава по некоторым критериям, например, по времени согласно формуле Бартона-Кеблера, или содержанием, используя методы Text Mining.

4.1. Агрегация информационных потоков

Одним из подходов к агрегированию информационных потоков можно считать создание информационно-поисковых систем, динамические базы данных которых аккумулируют документы из потоков. Выше были рассмотрены известные модели поиска информации, отмечены преимущества и недостатки отдельных методов поиска в применении их к информационным потокам. Для обеспечения эффективной навигации в современных информационных массивах, а также реализации поисковых процедур в режиме реального времени должны применяться современные средства автоматической группировки документов. Большое количество этих средств известно уже в течение десятилетий, но проблема их адаптации к динамическим информационным потокам остается открытой.

Существует много общего между существующими в настоящее время системами агрегирования информации. Однако, следует заметить, что многие из рассмотренных возможностей уникальны с точки зрения идеологии и технологии для каждой отдельной системы. Это, прежде всего, языки макроописания информационных ресурсов, которые позволяют интегрировать потоки слабо структурированных данных с сайтов сети Интернет, приведенных в различных форматах, элементы технологии Text Mining, в частности многоаспектные информационные портреты, динамические навигаторы, обзоры основных сюжетов, которые формируются в режиме реального времени, выявление новых событий и возможность построения таблиц взаимосвязи понятий.

4.2. Организация мониторинга и адаптивного агрегирования информации

Своевременное получение многоаспектной и объективной документальной информации с помощью средств мониторинга компьютерных сетей, современных поисковых и метапоисковых систем для последующего ее использования в научных исследованиях может быть достигнуто лишь путем внедрения новых теоретических и технологических решений. Поэтому особо актуальным является разработка теоретических и технологических принципов построения адаптивных информационных хранилищ, автоматизированных систем обработки и обобщения информации из документальных хранилищ сверхбольшого объема, которые должны стать основой для создания интеллектуальной среды решения аналитических междисциплинарных проблем.

Задачи мониторинга информационных потоков большого объема в компьютерных сетях, их адаптивного агрегирования и обобщения осложняются отсутствием типовых методик и решений, неполнотой существующих технологических подходов.

В результате проведенных исследований и обобщений предлагается модель системы мониторинга, адаптивного агрегирования и обобщения информации, функциональная схема которой приведена на рис. 65.

В соответствии с этой схемой, отдельный пользователь обращается к системе через специальный интерфейс. Предусматривается, что пользователь может быть зарегистрированным, тогда он имеет большие права, в частности, может управлять собственным информационным кэшем, чего не может делать незарегистрированный пользователь системы (посетитель). Кроме того, постоянные статические запросы отдельных пользователей могут вводиться для обработки в подсистеме выборочного распространения информации (средствами электронной почты) через администратора.

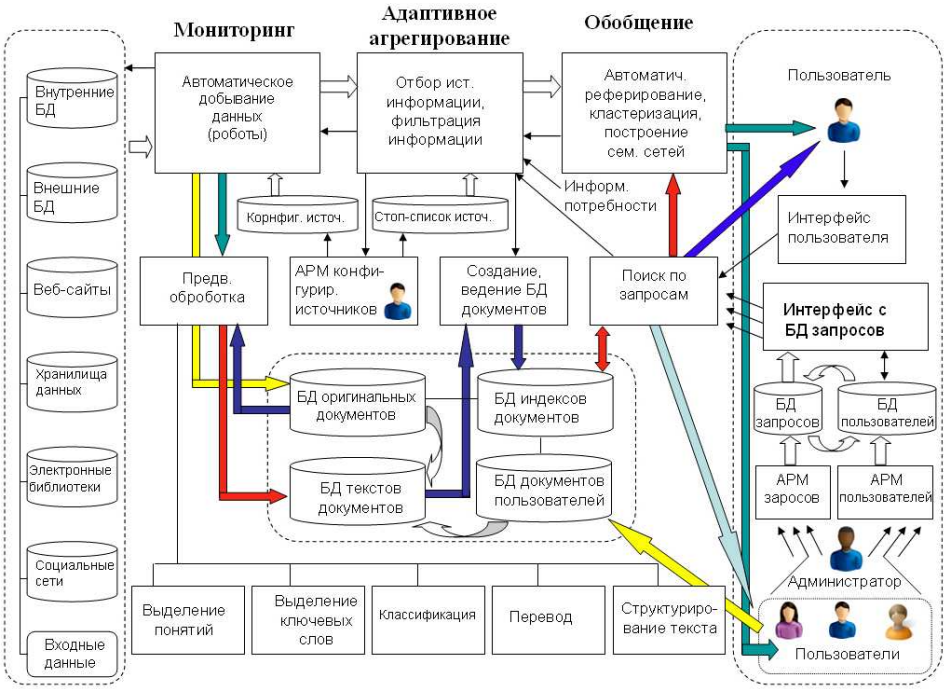


Рис. 65 – Функциональная схема системы мониторинга, адаптивного агрегирования и обобщения информации

Администратор вводит запросы пользователей с помощью специализированного автоматизированного рабочего места (АРМ). Для учета пользователей он также применяет отдельное программное обеспечение, с помощью которого ведется база данных пользователей. Содержимое базы данных запросов пользователей через специальный прикладной интерфейс передается метапоисковой системе, к которой также передаются запросы пользователей, которые работают с системой в режиме онлайн.

Все зарегистрированные пользователи имеют возможность размещать свои собственные документы в базу данных, к которой реализован доступ в поисковом режиме (также через метапоисковую систему). То есть ядром взаимодействия пользователя с системой является метапоисковая система, которая по существу агрегирует внешние и внутренние информационные ресурсы и предоставляет пользователю доступ к ним через "единственное окно".

Метапоисковая система обеспечивает взаимодействие с внешними сетевыми поисковыми системами, такими, как, например, Google, Bing, Rambler, с помощью которых реализуется поиск в таких информационных источниках, как веб-сайты, некоторые социальные сети, внешние базы данных, хранилища данных, архивы, электронные библиотеки и тому подобное.

Одним из главных принципов построения модели адаптивного документального хранилища является принцип фильтрации результатов. Он заключается в том, что во-первых, происходит фильтрация неинформативных сайтов или сайтов с недоступными первоисточниками (так называемый "черный список", или "стоп-список"). Кроме того, адаптивная метапоисковая система разбирает полученные результаты на отдельные документы и проверяет их доступность. Например, если на пути к документу присутствует доменное имя, присутствующее в "стоп-списке", то документ отбрасывается и не используется в

дальнейшей обработке. Это лишь один из критериев фильтрации. Те документы, которые прошли этап фильтрации, преобразуются для предоставления результатов пользователю. Также осуществляется поиск во внутренней базе этих файлов (в информационном кэше на прокси-сервере, который содержит найденные раньше документы). Если такие файлы были найдены, то выведение документа дополняется информацией о возможной доступности этого файла за найденной ссылкой. Если этот файл отсутствует по указанному адресу в Интернете, то выводится сообщение, что этот файл может быть отсутствующим. Если же информация об этом файле присутствует в информационном кэше и он вероятно существует, то вывод дополняется информацией, такой, как размер файла, а также создается HTML-версия этого файла. После подсчета количества найденных документов подготовленные результаты выводятся пользователю через стандартный веб-интерфейс.

Таким образом, пользователю предоставляются лишь те документы, которые прошли специальную фильтрацию. Фильтры создаются, с одной стороны, информационным администратором, который формирует "стоп-список" источников информации, доступ к которым нуждается в подписке, регистрации, содержит лишь метаданные относительно документов и так далее, а, с другой стороны, программными приложениями, которые не позволяют выдавать ссылки на несуществующие документы, или переправляют ссылку к кэшу системы, где целевые документы уже размещены в результате обработки предыдущих запросов.

Второй принцип построения адаптивного документального хранилища заключается в настройке на уже найденные пользователями документы. То есть реализуется модуль кэширования, основное задание которого – сбор ссылок на документы, которые получены в процессе работы с пользователем метапоисковой системы, чтобы в дальнейшем сохранить в информационном хранилище (кэше системы) файлы, а

также связанную с ними информацию, такую, как доступность файла по этой ссылке и размер файла.

Система периодически обновляет информацию о тех файлах, которые были сохранены в базе данных. Если файл не был раньше доступен, но доступный в тот момент, когда производится вторичное сканирование, информация в базе данных обновляется; если же он становится недоступным, то в базу данных записывается информация о недоступности этого файла, чтобы в дальнейшем предложить пользователю получить этот файл из кэша. Найденные и отмеченные пользователями как полезные документы подлежат предварительной обработке, в частности, они переводятся в текстовый формат, структурируются для отображения (по желанию пользователей) в этом формате. Кроме того, путем экстрагирования из документов выделяются отдельные понятия (персоны, компании, топонимы и тому подобное), по лингвостатистическим критериям выделяются ключевые слова, осуществляется перевод ключевых слов другими языками, классификация документов по тематикам. Выбранные документы загружаются к базы данных оригинальных документов, куда также (в необходимом объеме) загружаются в режиме автоматического мониторинга документы из выбранных администратором хранилищ данных, электронных библиотек, архивов. Каждый оригинальный документ сопровождается своим текстовым образом в базе данных текстовых документов. Отобранные понятия и ключевые слова загружаются в базу данных индексов документов, к которым и обращается внутренняя поисковая система.

В результате реализации функциональной схемы, которая рассматривается, пользователь по своим запросам может получать как перечни доступных релевантных документов, так и обобщенные отчеты, дайджесты, перечни сюжетных цепочек, интерактивные семантические сети, диаграммы трендов понятий или событий и тому подобное.

Основным критерием ранжирования информации в современной метапоисковой системе должен быть рейтинг поисковых систем. Так, например, у поисковой системы Google рейтинг более высок, чем у системы Bing (в Google больше охватывания ресурсов, более релевантные результаты). Если ссылка на один и тот же PDF-документ была получена из разных поисковых систем, то выбирается та из них, которая содержит наиболее полное описание.

Таким образом, предложенная обобщенная метапоисковая система реализует такие необходимые этапы обработки сетевой информации, как мониторинг, адаптивное агрегирование (по информационным потребностям пользователей) и обобщение информации, которая создает предпосылки для реализации поисковой среды нового типа, предназначенной для поддержки информационно-аналитической деятельности.

Приведенная архитектура системы значительно дополняет модель контент-мониторинга и создает условия ее эффективного использования за счет создания нового интерфейса.

4.3. Архитектура информационно-аналитической системы

Информационно-аналитическая система (ИАС) предназначена для аналитической обработки информации, данных, циркулирующих как в рамках корпоративной сети, так и поступающих из внешних источников. ИАС агрегирует, обеспечивает обработку и анализ информации, и ее хранение. Входящие в состав ИАС инструментальные средства обеспечивают преобразование больших объемов детализированных данных (Big Data) из хранилищ данных в обобщенную информацию, пригодную для принятия решений. Современная ИАС охватывает как представленные выше элементы агрегации и мониторинга данных, так и средства поддержки аналитической деятельности, создаваемые на основе концепции искусственного интеллекта, в том числе, экспертных систем, баз знаний,

онтологий предметных областей.

На рис. 66 представлена обобщенная функциональная схема современной ИАС. На вход такой системы поступает как внутенняя корпоративная информация, так и информация из внешних источников (прежде всего из Интернет). Информация может предоставляться по регламенту в виде отдельных порций или документов, так и в виде направленных информационных потоков, сканироваться с помощью специального программного инструментария, поступать в результате целенаправленного регулярного мониторинга или зондирования (фрагментального исследовательского сканирования) выбранных источников. Поступающая на вход системы информация обрабатывается специальными программами-конверторами, преобразуется к виду, пригодному для загрузки в хранилище данных, где сохраняется в виде баз данных или информационных массивов определенной структуры (например, в формате XML).

Информация из хранилища данных может быть доступной непосредственно через пользовательский интерфейс (в том числе и в поисковых режимах), интегрироваться в специальные приложения и выступать в качестве источника формализованных знаний, фактов добываемых с помощью специального инструментария экстрагирования и образующих в своей совокупности базу знаний ИАС.

Ведение базы знаний происходит в автоматизированном режиме. Администратору базы знаний и пользователям ИАС доступны также онтологии как структуры отображающие модель предметной области, которые формируются и сопровождаются с помощью специального инструментария – онтологического редактора. В качестве одного из источников для формирования онтологий выступает база знаний.

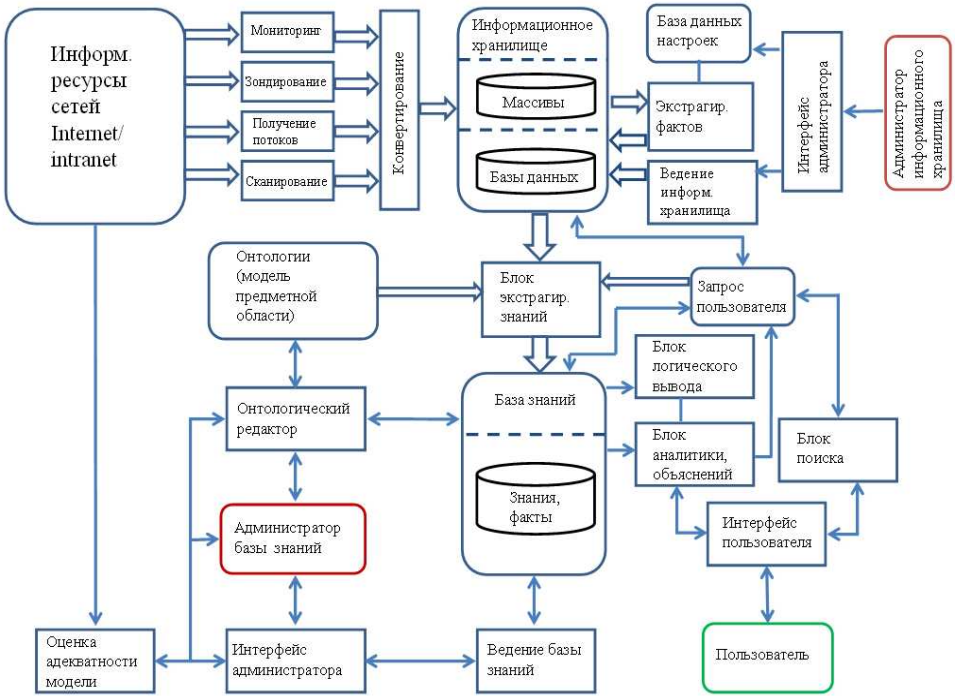


Рис. 66 – Функциональная схема информационно-аналитической системы

В процессе функционирования системы администратор базы знаний оценивает адекватность модели предметной области, являющейся семантической основой всей ИАС, сравнивая ее с потоком входных данных, поступающих в систему.

Таким образом, пользователю доступны не только информационная компонента системы, но и аналитическая поддержка, обеспечиваемая средствми онтологической навигации и логического вывода, интегрированными с базой знаний ИАС.

4.4. Корпоративная метапоисковая система Doc's Bundle

В настоящее время ни одна из традиционных поисковых систем на достаточном уровне не помогает при поиске актуальной документальной информации, которая находится в динамической части сети Интернет. Решение этой задачи требует применения системы-посредника между пользователем и сетью. Подобный посредник должен выполнять работу по сбору, селекции информации и осуществлять предварительную обработку данных для создания документального информационного хранилища.

4.4.1. Принципы построения корпоративной метапоисковой системы

В настоящее время в интернет-пространстве содержится большое количество документальных ресурсов, представленных в формате PDF [Document, 2008]. Популярность данного формата вызвана тем, что он является компактным и удобным для хранения информации, представленной изначально в различных видах: простого текста, векторных и растровых изображений, страниц веб-сайтов, форм и мультимедийных файлов. Вместе с тем, при поиске необходимой документации в формате PDF с помощью традиционных сетевых информационно-поисковых систем пользователь постоянно сталкивается с проблемами, связанными с плохой доступностью целевой информации (условиями платного доступа,

отсутствием необходимых файлов по указанным адресам, или неверными гиперссылками). Хотя большинство поисковых систем, таких как Google, Yandex, Rambler, Yahoo и пр. выводят в список результатов информацию о найденных PDF-файлах, вместе с тем они часто дают ссылки на несуществующие PDF-файлы, или ссылки на веб-сайты, где PDF-файлы находятся в закрытом доступе. Например, указывая в строке адреса название PDF-файла (полученное с помощью Google Scholar) 36W622113036P357.pdf на сервере такого популярного издания, как Springer, пользователь получает не искомый документ, а его описание и регистрационную форму. Сказанное относится и к специализированным поисковым системам, ориентированным на поиск документов в формате PDF (например, OSUN – www.osun.org, PDFGod, www.pdfgod. com, pdf-search-engine.com/ и др.) В указанных поисковых системах нет возможности отсортировать или отфильтровать результаты поиска или просто поискать в базе данных с уже сохраненными PDF-документами. Все перечисленные системы поиска PDF-документов основаны на поиске информации в других поисковых системах. В основном они направлены на англоязычный сегмент пользователей, и используют для получения информации в основном систему Google, что ограничивает выдаваемые результаты. Кроме того, лишь одна из специализированных поисковых систем может выдавать PDF-файлы в HTML-виде (это удобно для оперативного ознакомления с содержанием документов) – это pdf-search-engine.com.

Ниже будет рассмотрена технология агрегирования документальных информационных потоков, реализованная в виде метапоисковой системы Docs Bundle, доступной в настоящее время по адресу http://docsbundle.info.

Как прототип обобщенной системы мониторинга, адаптивного агрегирования и обобщения информационных потоков авторами предложено модельное решение – система Doc's Bundle, которая

позволяет искать документы в формате PDF как в Интернете, так и в специально накопленном кэше документов (внутри системы) в процессе работы. Формат PDF как основной для модельного решения был выбран потому, что в настоящее время в интернет-пространстве находится большое количество документальных ресурсов, представленных в этом формате. Популярность этого формата обусловлена тем, что он является компактным и удобным для хранения информации, представленной с самого начала в виде простого текста, векторных и растровых изображений, страниц веб-сайтов, форм и мультимедийных файлов.

Любая поисковая система в процессе работы просматривает определенный набор серверов и отбирает документы в соответствии с заданными критериями. Сегодня поиск с помощью разных систем по одним и тем же ключевым словам дает различные результаты. Это привело к идее создания так называемых метапоисковых (или мультипоисковых) систем [Meng, 2002], которые обращаются за помощью сразу к нескольким поисковым системам. Каждая из метапоисковых систем имеет свой язык запросов. Метапоисковая система переводит сформулированный на ее языке запрос на языки, используемые каждой машиной поиска. Далее, результаты поиска всеми системами объединяются и представляются в соответствующей форме. Естественно, поиск с помощью метапоисковых систем занимает больше времени по сравнению с обычными ИПС.

С помощью метапоисковой системы Docs Bundle можно искать PDF-файлы в таких поисковых системах как Google, Bing, Яndex, Rambler, а также в ее собственной базе данных (кэше Docs Bundle). Поиск в кэше производится при любом запросе по умолчанию и выводится списком ниже результатов полученных от других ИПС.

Особенностью Docs Bundle является то, что она полностью направлена на поиск доступных пользователю PDF-файлов, с возможностью фильтрации

платных ресурсов, текстовых описаний, любой информации, кроме самих файлов, без сопровождающего их информационного шума или рекламы.

Общая схема работы местапоисковой системы Docs Bundle охватывает ряд этапов. После того, как пользователь задает запрос метапоисковой системе, с ее помощью создаются запросы для каждой поисковой системы, учитывая уникальные возможности их синтаксиса. Затем модифицированные запросы пересылаются поисковым системам, которые возвращают результаты поиска. После этого метапоисковая система разбирает полученные результаты на отдельные документы и проверяет их доступность. Например, если в пути к документу присутствует доменное имя, присутствующее в стоп-списке, то документ отбрасывается и не используется в дальнейшей обработке. Это лишь один из критериев фильтрации. Те документы, которые прошли этап фильтрации преобразуются для вывода результатов пользователю. Также производится поиск во внутренней базе данных файлов (в информационном кэше на прокси-сервере, содержащем найденные ранее документы [Додонов, 2006]). Если такие файлы были найдены, то вывод документа дополняется информацией о возможной доступности этого файла по обнаруженной ссылке. Если данный файл отсутствует по указанному адресу в Интернет, то выводится сообщение о том, что данный файл может отсутствовать. Если же информация о данном файле присутствует в информационном кэше и он предположительно существует, то вывод дополняется информацией такой как размер файла, а также создается HTML версия этого файла. После подсчета количества найденных документов подготовленные результаты выводятся пользователю через стандартный веб-интерфейс [Ландэ, 2010].

Таким образом, система Docs Bundle состоит из трех основных модулей (рис. 67):

- метапоисковая система;
- модуль кэширования информации (информационный прокси-сервер);
- внутренняя поисковая система, работающая как с информационным прокси-сервером, так и репозиторием.

Рис. 67 – Модель системы адаптивного агрегирования информации

Основным критерием ранжирования информации в системе Docs Bundle является рейтинг поисковых систем. Так, например, у поисковой системы Google рейтинг выше, чем у системы Bing (в Google больший охват ресурсов, более релевантные результаты). В Docs Bundle происходит фильтрация неинформативных сайтов или сайтов с недоступными первоисточниками (так называемый «черный список»).

Если ссылка на один и тот же PDF-документ была получена от нескольких поисковых систем, то выбирается та из них, которая содержит более полное описание.

Результаты представляются пользователю в виде списков результатов нескольких поисковых систем, которые следуют друг за другом.

В системе Docs Bundle используется модуль кэширования, основная задача которого – сбор ссылок

на PDF-документы, которые получены в процессе работы с пользователем метапоисковой системы, чтобы в дальнейшем сохранить в информационном хранилище (кэше Docs Bundle) файлы, а также сопутствующую им информацию, такую как: доступность файла по данной ссылке, размер файла.

Поиск в кэше Docs Bundle и ранжирование полученных результатов происходит по иному принципу. Так как в системе уже загружены тексты pdf-файлов, то строятся собственные таблицы релевантности с учетом частоты встречаемости ключевых слов, их позиции (если ключевое слово встречается в названии, то данный документ более релевантен чем тот, в котором ключевое слово встречается в середине текста).

Система периодически обновляет информацию о тех файлах, которые были сохранены в базе данных Docs Bundle. Если файл не был ранее доступен, но доступен в тот момент когда производится вторичное сканирование, информация в базе данных Docs Bundle обновляется, если же он становится недоступным, то в базу данных записывается информация о недоступности данного файла, чтобы в дальнейшем предложить пользователю получить этот файл из кэша. Далее PDF-файл кэшируется, конвертируется в текст, затем строится поисковый индекс этого файла.

Во внутреннем формате для каждого файла присутствует такая информация как текстовый вариант PDF-файла, размер файла, ссылка, по которой был сохранен файл, ссылки на похожие файлы с других сайтов.

Внутренняя информационно-поисковая система позволяет пользователю искать в кэше системы Docs Bundle документы, которые динамически накапливаются. Каждый документ во внутренней поисковой системе ранжируется по релевантности. Критериями релевантности документа являются: количество вхождений ключевых слов (по которым пользователь ищет документ), размер документа, а

также наличие подобных документов в базе данных метапоисковой системы. Результатом поиска информации в кэше Docs Bundle является аннотированный список найденных документов. Аннотации (сниппеты) документов – строки с первыми вхождениями ключевых слов, введенных пользователем.

Метапоисковая система Docs Bundle изначально была создана как система метапоиска научно-технической документации и использовалась пользователями, которые искали именно такие документы. Соответственно в адаптивном кэше Docs Bundle присутствуют преимущественно научно-технические документы (их количество превышает 1 млн.) из более чем 200 тысяч источников.

Лидируют среди источников для системы Docs Bundle сайты nbuv.gov.ua (Национальная библиотека Украины им. В.И. Вернадского), ioffe.ru (Физико-технический институт имени А.Ф. Иоффе), window.edu.ru (Единое окно, доступ к образовательным ресурсам) и др.

Именно благодаря эффекту адаптивности, наличию большого количества информации, уже загруженной с научных сайтов, журналов, серверов препринтов и т.д., можно констатировать, что сегодня система Docs Bundle лучше всего настроена на поиск научно-технической информации.

Сравнение результатов эксплуатации системы Docs Bundle с другими подобными системами позволяет сделать заключение не только о том, что эта система лучше отфильтровывает недоступные пользователю документы, но и о ее лучшей ориентации на русский язык.

Рассмотренная модель, реализованная в виде метапоисковой системы Docs Bundle, уже в настоящее время нашла своих пользователей и позволила сформулировать более сложные задачи, которые должны быть решены в рамках отдельной научно-исследовательской работы.

4.4.2. Интерфейс пользователя корпоративной метапоисковой системы

Во время выполнения поиска пользователь системы Doc's Bundle попадает на страницу (рис. 68), где представлены результаты поиска из нескольких поисковых систем, которые были выбраны им в списке. Также пользователю предлагается список ключевых слов для быстрой навигации в системе. Документ в списке представлен такими данными:

- название документа со ссылкой на источник;
- аннотация документа;
- ссылка на скачивание документа из кэша системы, если данный документ попал в кэш, размер файла;
- вспомогательная информация о том, в какой поисковой системе документ был найден.

Для того, чтобы пользователь использовал максимальные возможности системы, он должен войти в систему или зарегистрироваться. Для входа в систему пользователь должен ввести логин и пароль, после чего ему предоставляются расширенные возможности:

- загрузка собственных файлов в кэш для хранения и поиска в нем;
- отбор файлов в "Избранное", а также поиск в "Избранном";
- создание собственных рубрик.

На странице поиска документов пользователю доступно:

- меню пользователя, которое позволяет использовать расширенные функции системы;
- форма поиска с дополнительным параметром, с помощью которого можно проводить поиск в "Избранном";

- список документов;
- блок пользовательских рубрик;
- блок ключевых слов для навигации пользователя по системе.

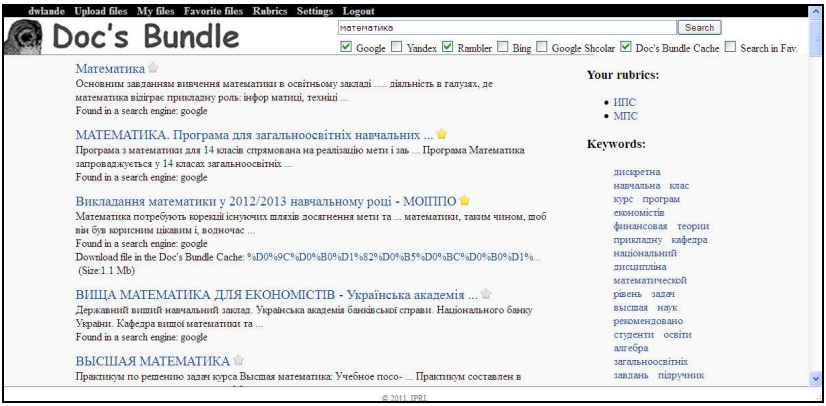


Рис. 68 – Страница поиска PDF-документов

Документ в списке представлен такими данными:

- название документа со ссылкой на источник;
- специальный инструмент "Звездочка", с помощью которого пользователь может прибавить/удалить документ в "Избранное";
- аннотация документа;
- ссылка на скачивание документа из кэша системы, если данный документ попал в кэш, размер файла;
- вспомогательная информация относительно того, в какой информационно-поисковой системе данный документ был найден.

Зарегистрированный пользователь может загружать в систему свои документы в формате PDF, а также проводить поиск в закачанных файлах.

Пользовательские файлы принимают участие в поиске и для других пользователей, в случае если был

выбран поиск в кэше системы.

Файлы, добавленные пользователем в систему, автоматически включаются в список выбранных документов. В случае, если пользователь снял из файла отметку "Избранный" файл спустя некоторое время может быть удален.

Выбранные файлы – дополнительная функциональность, с помощью которой пользователь может отложить полезный для него файл и в дальнейшем не проводить дополнительный поиск.

Чтобы добавить или удалить документ в "Избранное" пользователь должен нажать на специальный элемент интерфейса "Звездочка". Если звездочка окрашена в желтый цвет – документ добавлен к избранному, в серый – документ не в избранном.

Пользователь может пересмотреть весь список избранных файлов на странице Favorite files или поискать по запросу в "Избранном" отметив в поисковой форме параметр Search in Fav.

Все файлы, добавленные в "Избранное", не удаляются из кэша системы, и существуют до тех пор, пока существует хотя бы один пользователь, который добавил файл в "Избранное". Кроме того, зарегистрированный пользователь может добавлять, изменять и удалять рубрики (рис. 69) пользователей. На странице поиска пользователю представлен список его рубрик в виде отдельного блока.

Рассмотренная модель уже в настоящее время нашла своих пользователей и позволила сформулировать более сложные задачи, которые должны быть решены при построении корпоративных информационно-аналитических систем.

Рис. 69 – Страница добавления рубрик пользователей

Предусматривается, что результаты проведенных исследований составят теоретическую базу для разработки автоматизированных систем мониторинга, адаптивной агрегации и обобщения информационных документальных потоков, построения и ведения информационных ресурсов сверхбольших объемов и разнообразной тематической направленности, позволят соединить в единственной технологической цепочке мониторинг, информационный поиск, агрегацию информации с содержательным анализом данных, их обобщением, которое повысит качество обработки информации из глобальных сетей, и, соответственно, эффективность информационно-аналитической поддержки научно-аналитической деятельности отечественных ученых и специалистов.

5. ИНФОРМАЦИОННЫЕ ОПЕРАЦИИ

В последние годы благодаря многочисленным документам и публикациям Министерства обороны США стал популярен термин «информационные операции», прежде всего потому, что информационные технологии играют постоянно увеличивающуюся роль в военных операциях. При этом информационные операции определяются как «акции, направленные на воздействие на информацию и информационные системы противника, и защиту собственной информации и информационных систем» [DoD, 2003]. Информационные операции рассматриваются как объединение основных возможностей радиоэлектронной войны, компьютерных сетевых операций, психологических операций, военных действий и операций по обеспечению безопасности с целью воздействовать, разрушать, искажать информацию, необходимую для принятия противником решений, а также защищать собственную информацию.

Информационные операции охватывают целый комплекс процессов, проводимых в самых разных областях. При этом необходимо отметить, что информационные операции – существенная и традиционная составляющая боевых операций. Несмотря на то, что формальное определение в документах Департамента обороны США ориентировано на военные аспекты информационных операций, оно вполне применимо практически для любой области жизнедеятельности.

Ниже будут рассматриваться такие информационные операции, которые реализуются с помощью информационных систем. Живучесть этих ИС во многом определяет живучесть информационных операций, которые реализуются в виде информационных воздействий на сознание людей.

Информация является отражением вложенного в нее смысла, поэтому сегодня информация превратилась из абстрактного термина в объект, цель и средство информационных операций, стала критическим

понятием в проблематике безопасности. Бывший министр обороны США Уильям Коэн 18 марта 1999 г. заявил, что «способность армии использовать информацию, чтобы доминировать в будущих сражениях, даст США новый ключ к победам в течение многих лет, если не в течение нескольких поколений» [Hill, 2000].

При моделировании и проведении информационных операций необходимо учитывать значение ценности информации для лиц, принимающих решение (ЛПР). Ценность информации включает ее своевременность, точность и «аналитичность». С практической точки зрения ценность информации также может быть определена как ее значимость или применимость, пригодность к использованию. Под применимостью информации понимается обеспечение доступа ЛПР к готовой для использования информации. Стандарт ISO 9241 (ISO – Международная Организация по Стандартизации) определяет применимость в терминах эффективности и удовлетворения потребностей указанного набора пользователей для решения указанного набора задач в специфическом окружении. На практике большая часть полезной информации поступает к ЛПР от информационно-аналитических систем, обеспечивающих ориентацию в ситуации и поддержку при принятии решений. Согласно полевому уставу военного ведомства США «Информационные операции» (FM 100-6), «ориентация в ситуации означает комбинацию ясного представления о диспозиции своих и вражеских сил с оценкой ситуации и намерений со стороны командования».

Информационные операции осуществляются в некоторой социальной среде, соответственно, для успешного их проведения необходимо адаптироваться к этой среде, преодолеть определенный барьер не очень сильного внимания к информационному воздействию. Этот барьер возникает благодаря так называемой иммунной системе среды, которая может не пропустить информационные воздействия, если она достаточно мощная и/или уже научилась защищаться от подобных

воздействий. К подготовительным действиям для проведения информационных операций может относиться создание «иммунодефицита» социальной среды путем воздействия через информационное пространство, например, с помощью материалов в СМИ. Очень часто информационные воздействия используют механизмы «вирусного маркетинга», например, в виде слухов, когда сенсационно поданная дезинформация распространяется с огромной скоростью. Именно иммунная система оказывает противодействие подобным информационным операциям. Очень часто с иммунной системой общества отождествляют государство, призванное обеспечивать безопасность этого общества, т.е. при наличии сильного государственного аппарата вероятность успеха антиобщественных информационных операций существенно снижается. Читатель прекрасно знает, как происходило противодействие подобным информационным процессам в тоталитарных государствах. В демократическом обществе, естественно, тоталитарные методы не применимы. В этом случае иммунитет достигается за счет «обучения», т.е. демократическое общество должно пройти через многие информационные атаки, воздействия, влияния стереотипов, чтобы выработать необходимый иммунитет.

Уровень готовности к проведению информационных операций сегодня считается ключевым фактором успеха проведения любой социальной процедуры, кампании.

Особой целью при проведении информационных операций являются информационно-аналитические системы субъекта воздействия. Оказывая влияния на такие системы, можно добиться того, что принимающие решение лица из лагеря противника примут неадекватные выводы, и требуемый социальный процесс изменит траекторию в необходимом оказывающей влияние стороне направлении [Горбулін, 2009] (рис. 70).

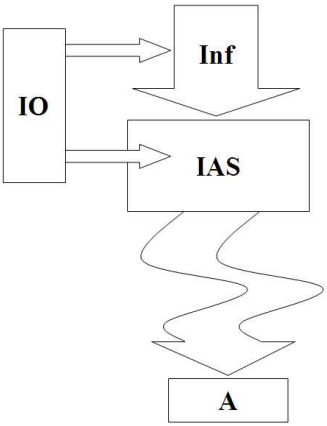


Рис. 70 – Воздействие на информационно-аналитическую систему противника: Inf – информационное пространство; IAS – информационно-аналитическая система; А – абонент системы – ЛПР; IO – информационные воздействия

В данном случае к непосредственным информационным воздействиям может быть отнесено размещение в информационном пространстве документов, компрометирующих противоположную сторону, реклама (в том числе скрытая) своих преимуществ, искаженные данные о внешней среде, искаженная информация о намерениях и т.д.

Социальные процедуры и процессы, как правило, сложно оценивать и моделировать, так как их результаты относятся к психологическим и социологическим, а не физическим. Именно этот факт также определяет проблематичность прогнозирования результатов моделирования информационных операций. Кроме того, экспериментирования с информационными воздействиями в рамках информационных операций более сложны и опасны, чем при моделировании физических процессов.

Для достижения эффективности влияния на процессы принятия решения противником иногда необходимо предпринимать действия в течение

длительного времени, прежде чем они вступят в силу.

Одна из основных компонент информационных операций – социальное влияние, охватывающее все многообразие процессов влияния. Существенные изменения в убеждениях или отношении людей к некоторой проблеме или явлению, как ожидается, будут вести к изменению в поведении, связанным с этой проблемой.

В 1948 году Харольд Д. Лассвел [Lasswell, 1948] разработал модель трансмиссии коммуникаций, состоящую из пяти компонент:

- источник – персона, которая влияет или убеждает другие персоны;
- сообщение – с помощью чего источник пробует убедить цель;
- цель – человек, на которого источник пробует влиять;
- канал – метод доставки сообщений;
- воздействие – реакция цели на сообщение.

Хотя Лассвел прежде всего интересовался массовой коммуникацией, его модель передачи информации может применяться в межличностной коммуникации типа циркулярных моделей Шеннона–Вивера (Shannon–Weaver) и Осгуда–Шрамма (Osgood–Schramm), которые включают петли обратной связи в процессе коммуникаций, утверждая, что коммуникация является циркулярным, а не линейным процессом [Schramm, 1974], [Osgood, 1954].

Моделирование объективных факторов социального влияния требует междисциплинарных подходов, имеющих отношение к информатике, маркетингу, политологии, социальной психологии. Самые известные модели формирования общественного мнения и социального влияния базируются на теории Латэйна динамического социального воздействия [Latane, 1981], [Latane, 1997], развитой многими другими авторами,

прежде всего, в работах [Nowak, 1990], [Lewenstein, 1993], [Kacperski, 2000], [Sobkowicz,2003].

Пытаясь обосновать механизм социального влияния сообщений Латэйн [Latane, 1981] подчеркнул важность трех признаков отношений источника и цели:

- сила – социальная сила, вероятность или уровень влияния на индивидуумов;
- непосредственность – физическое или психологическое расстояние между индивидуумами;
- число источников – количество источников, стремящихся к цели.

Современное состояние моделирования информационных операций характеризуется рядом открытых проблем, основные из которых относятся к пониманию понятий информационного влияния и воздействия.

5.1. Информационное влияние

Универсальными характеристиками объектов являются его состояние и возможность воздействия на другие объекты. Реализация возможности воздействия требует определенных условий, которые принято называть его влиянием. При этом объект, который может осуществлять свою волю, называют субъектом, а управлением принято называть воздействие по отношению к объекту воздействия, применяемое с определенной целью.

Когда индивидуум является целью влияния одного или более источников, динамическая социальная теория воздействия утверждает, что уровень социального влияния на индивидуума может быть представлен уравнением, являющимся основой так называемой индивидуум-ориентированной модели:

$$I_i = -S_i\beta - \sum_{j=1,\, j\neq i}^{N} \frac{S_j O_j O_i}{d_{i,j}^{\alpha}},$$

где I_i – величина (количество) социального давления, оказываемого на индивидуума i, $(-\infty < I_i < \infty)$; O_i и O_j представляет мнение индивидуума (i и j, соответственно) по актуальному вопросу – +1 или –1 – поддержку или возражение относительно данного вопроса, соответственно. S_i (S_j) представляет силу индивида i (j) или влияние ($S_i > 0,\ S_j > 0$); β – сопротивление индивидуума к изменениям ($\beta > 0$); $d_{i,j}^{\alpha}$ – расстояние между индивидуумами i и j ($d_{i,j}^{\alpha} \geq 1$); α – показатель сокращения расстояния ($\alpha \geq 2$); N – общее количество агентов (индивидуумов, составляющих сообщество). Значение β, тенденция сохранять собственное мнение или сопротивляться изменению определяет то, что индивидуумы в рамках модели могут требовать больших или меньших объемов социального давления для изменения их мнения. Большие уровни значения α соответствуют эффекту возрастания расстояния между источником и целью, что влияет на объем социального давления на цель.

На основе введенных терминов формулируется понятие «информационного поля объекта» [Кононов, 2003], описываются его характеристики. Это дает возможность определить информационное воздействие как воздействие на информационное поле объекта. Исследуя информационные поля объектов и субъектов социальных систем, можно определить информационные влияния и управления. При этом информация может рассматриваться и как объект, и как средство воздействия. Использование информации как средства воздействия требует в процессе управления осуществить подготовку данных, производство соответствующей информации, а лишь затем реализовывать созданную информацию в виде воздействия (влияния).

Одним из основных методов ведения

информационных операций является информационное влияние, оказываемое с целью информационного управления. Под информационным управлением в данном случае понимается механизм управления, когда управляющее воздействие носит неявный, косвенный информационный характер и объекту управления дается определенная информационная картина, под влиянием которой он формирует линию своего поведения. Таким образом, информационное управление — это способ воздействия, побуждающий людей к упорядоченному поведению, выполнению требуемых действий.

В соответствии с [Кононов, 2003], [Кульба, 2004] процесс информационного влияния одного объекта на другие целесообразно декомпозировать на следующие этапы:

- генерация источником влияния данных, информационных элементов и информационных совокупностей;
- передача информации источником влияния;
- прием информации реципиентом;
- генерация совокупности данных, информационных элементов и новых совокупностей объекта влияния;
- соответствующие активные действия объекта влияния.

Информационные воздействия на элементы систем можно классифицировать по таким признакам, как источники возникновения, длительность воздействия, природа возникновения и т.п.

Для выбора конкретных способов реализации информационного управления необходимо конкретизировать задачи, решаемые с помощью информационного воздействия, провести анализ процесса формирования информационных операций и выработать критерии их оценки. Информационное

управление рассматривают как процесс, охватывающий такие три взаимосвязанных направления:

- управление обменом данными между реальным миром и виртуальным миром субъекта влияния;
- управление виртуальным миром субъектов влияния, механизмами принятия решений;
- управление процессом преобразования решений в действия субъектом влияния в реальном мире.

Информационное воздействие может быть двух основных видов:

1) изменение в требуемую сторону данных, которые использует информационно-аналитическая система объекта воздействия при принятии решений;

2) непосредственное влияние на процесс принятия решения объекта воздействия, например, на процедуры принятия решения или отдельные лица, принимающие решения.

Важнейшее значение для проведения информационных операций имеет окружающая среда, состояние объектов информационного воздействия, их взаимное влияние. В частности, если в качестве объектов информационных операций выбирается некоторое электоральное поле, то важно учитывать все электоральные популяции, входящие в это поле, которые представляют сторонников (или противников) тех или иных политических сил. Несмотря на то, что в дальнейшем будут рассматриваться и некоторые модели, в которых в явном виде постулируется однородность среды, в общем случае по отношению к информационным операциям окружающая среда может состоять из областей:

- доминирующего восприятия;
- повышенной чувствительности;
- индифферентности к соответствующим информационным воздействиям.

5.2. Этапность информационных операций

Остановимся отдельно на этапности информационных операций. Очевидно, не существует единственного «стандартного» плана проведения как наступательных, так и оборонительных информационных операций. Можно лишь рассмотреть примерную, полученную путем обобщения некоторых уже реализованных информационных операций последовательность действий при их осуществлении.

На практике информационная операция как процесс информационного воздействия на массовое сознание, как правило, реализуется следующим образом: в результате предварительной разведки вырабатывается план следующего этапа — оперативного управления и намечаются соответствующие мероприятия оперативной разведки, которые являются приближенной моделью решения, после чего реализуется оперативное управление противником. На этапе оперативной разведки определяется уровень отклонения первоначальной модели от реальности, и если оно незначительно, то реализуется первоначальный план. В противном случае строится новый план оперативного управления и управления противником. Далее цикл повторяется до тех пор, пока оперативная разведка не подтвердит используемую модель. При этом окончательное решение принимается с определенным оперативным риском.

Таким образом, процесс информационного воздействия охватывает такие основные этапы [Чхартишвили, 2004] (рис. 71):

- предварительная разведка (preliminary intelligence, PI);
- выявление текущей обстановки, состояния противника (Op);
- управление противником (management of enemy, M) (информационное воздействие на противника с целью передачи ему сведений

соответствующих замыслу управляющего);

- оперативная разведка (operational intelligence, OI) (проверка результатов рефлексивного управления);
- оперативное управление (operational management, OM) – действия управляющего для достижения требуемой цели.

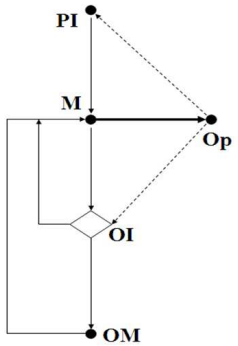


Рис. 71 – Основные этапы информационных операций

При планировании или моделировании социальных процессов, в частности информационных операций, всегда необходимо учитывать, что общее поведение социальных систем невозможно определить, оперируя исключительно рафинированными математическими моделями. Это главным образом обусловлено тем, что такие процессы в большой степени зависят от социально-психологических факторов.

Различают два основных типа информационных операций — наступательные и оборонительные. Однако, на практике, большая часть информационных операций является смешанной. Кроме того, большинство процедур информационных операций относятся одновременно к наступательным и оборонительным. Каждый из типов информационных операций, включая приведенные выше основные этапы, подразумевает некоторые особенности и уточнения.

Особенностью наступательных информационных операций (информационных атак) является то, что объекты воздействия таких операций определены и планирование основывается на достаточно точной информации об этих объектах. Информационная атака чаще всего требует нахождения или создания информационного повода (для оборонительных информационных операций поводом может являться сама информационная атака противника), раскрутка этого повода, т.е. пропаганда (в отличие от мер контрпропаганды при оборонительных информационных операциях), а также необходимость принятия мер по препятствию информационному противодействию.

Таким образом, план типовой информационной операции включает совпадающие на верхнем уровне для информационных операций обеих типов такие этапы, как оценка, планирование, исполнение и завершающая фаза. Приведем более детальный перечень компонент информационных операций.

В наступательных информационных операциях можно выделить такие основные фазы:

1. Оценка необходимости проведения операции:

    1) определение цели, прогноз достижимости, степени влияния;

    2) сбор информации.

2. Планирование.

3. Исполнение информационного воздействия:

    1) нахождение или создание информационного повода;

    2) раскрутка информационного повода (пропаганда);

    3) оперативная разведка;

    4) оценка воздействия;

5) препятствие информационному противодействию;

6) корректировка информационного воздействия.

4. Завершающая фаза:

1) анализ эффективности;

2) использование позитивных результатов информационного воздействия;

3) противодействие отрицательным результатам.

Типовая оборонительная информационная операция охватывает такие основные этапы:

1. Оценка:

1) анализ возможных уязвимостей (целей);

2) сбор информации о возможных операциях;

3) определение возможных «заказчиков» информационных воздействий:

– определение сфер общих интересов объекта и потенциальных «заказчиков»;

– ранжирование потенциальных заказчиков по их интересам.

2. Планирование:

1) стратегическое планирование оборонительной операции (явное или неявное):

– определение критериев информационных воздействий;

– моделирование информационных воздействий с учетом: связей объекта; динамики воздействия; «особых» (критичных) точек воздействия;

– прогнозирование следующих шагов;

– расчет последствий.

2) тактическое планирование контроопераций.

3. Исполнение – отражение информационного воздействия:

1) выявление и «сглаживание» информационного повода;

2) контрпропаганда;

3) оперативная разведка;

4) оценка информационной среды;

5) корректировка информационного противодействия.

4. Завершающая фаза:

1) анализ эффективности;

2) использование позитивных результатов информационного воздействия;

3) противодействие отрицательным результатам.

Оперативное управление информационными операциями с использованием информационно-аналитических систем можно проиллюстрировать с помощью диаграммы, представленной на рис. 72.

В соответствии с приведенной диаграммой информация из реального мира (R) поступает в информационное пространство, в частности, в средства массовой информации (I) либо непосредственно экспертам (E), также через средства массовой информации. От экспертов или непосредственно из информационного пространства (например, с помощью средств контент-мониторинга) информация поступает в информационно-аналитическую систему (IAS).

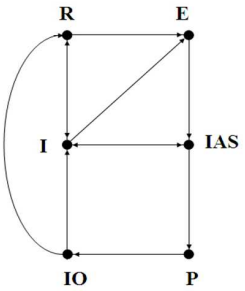


Рис. 72 – Диаграмма оперативного управления с использованием информационно-аналитических систем

Информационно-аналитическая система передает лицам, принимающим решения (Р), данные, которые определяют меры информационного воздействия на информационное пространство и непосредственно на объекты реального мира (людей, окружающую среду, компьютерные системы и т. д.)

5.3. Моделирование информационных операций

Моделирование можно рассматривать как один из способов решения проблем, возникающих в реальном мире, в частности, при планировании и проведении информационных операций. Чаще всего моделирование применяется в случаях, если эксперименты с реальными объектами невозможны, трудоемки либо слишком затратные. Моделирование охватывает отображение реальной проблемы в мир абстракции, изучение, анализ и оптимизацию модели, и отображение оптимального решения обратно в реальный мир.

При моделировании существуют два альтернативных подхода – аналитическое и имитационное моделирование. Идеальные аналитические модели допускают строгое аналитическое решение или, по меньшей мере, постановку, например в виде систем дифференциальных уравнений. Однако, аналитические решения не всегда реализуемы. Поэтому, особенно в последнее время, и особенно при решении

задач из области социальной динамики все чаще применяются методы имитационного моделирования (*Simulation Modeling*). Имитационное моделирование представляет собой более мощное и практически незаменимое средство анализа социальных процедур. Имитационную модель можно рассматривать как множество правил, определяющих будущее состояние системы на основании текущего. При этом процесс моделирования заключается в наблюдении эволюции системы во времени по данным правилам, и, соответственно, оценки адекватности модели, когда это возможно.

Наиболее перспективным направлением моделирования информационных операций является математическое описание самоорганизации среды восприятия и распространения информации с учетом сложившихся в текущий момент условий. Самоорганизующиеся среды, для которых отсутствует центральный механизм управления, а развитие идет за счет множества локальных взаимодействий, изучаются теорией сложных систем. Эта теория охватывает такие отрасли знаний, как нелинейная физика, термодинамика неравновесных процессов, теория динамических систем. Взаимодействие между отдельными элементами сложных систем определяет возникновение сложного поведения при отсутствии централизованного управления. Для исследований подобного поведения применяются самые современные методы, которые охватываются междисциплинарной основой современной методологии – концепцией сложности. В настоящее время к теоретическим и технологическим основам этой концепции относятся теории детерминированного хаоса, фракталов и сложных сетей, синергетика, волновой (вейвлет) анализ, многоагентное моделирование, теория самоорганизованной критичности (изучающей динамическое развитие до критического состояния, характеризуемого сильными пространственно-временными флуктуациями, без внешнего управления [Bak, 1996]), теория перколяции (Percolation –

протекание) и т.п.

Моделирование социальных процедур (информационные операции, безусловно, относятся к таковым) предполагает проведение вычислительных экспериментов, так как чаще всего возникают существенные ограничения, затрудняющие проведение «полевых» натурных экспериментов.

При моделировании информационных операций вычислительный эксперимент позволяет сократить операции по уточнению ограничений, подбору исходных данных, выбору правил функционирования компонент модели и т.д. В этом случае появляется возможность учета случаев, трудно реализуемых на практике, используя реальные данные лишь для идентификации параметров математической модели. Вместе с тем математическое моделирование имеет свои ограничения, реальный мир оказывается сложным для моделирования с достаточным уровнем детализации и точности, т.е. более или менее достоверные математические модели настолько сложны и многопараметричны, что не поддаются анализу и оценкам точными методами.

Отработать математические модели при планировании информационных операций можно лишь в процессе моделирования конкретных процедур, постоянно сопоставляя их результаты с реальностью.

Выраженная цель методологии оценки информационных операций состоит в том, чтобы обеспечить своевременный и точный анализ возможных несоответствий между запланированной операцией и фактическим воздействием. Когда обнаруживаются существенные различия, которые влияют на вероятности успеха операции, аналитическая система должна сообщать об этом лицам, принимающим решения, для того, чтобы откорректировать текущие планы и решения. Вместе с тем, при планировании информационных операций нельзя действовать методом проб и ошибок, поэтому необходимо развивать методы, позволяющие обобщать ретроспективные данные, и на

их основе проверять адекватность моделей.

В основу успешных моделей информационных операций закладываются синергетические подходы. Действительно, общество является сложной системой, каждая компонента которой характеризуется множеством признаков, имеет множество степеней свободы. При этом важным свойством этой системы является самоорганизация, которая является результатом взаимодействия таких компонент, как случайность, многократность, положительная и отрицательная обратная связь.

Особенностью математического моделирования информационных операций следует считать сравнительную простоту интерпретации получаемых результатов. Такие понятия, как «численность электората», «политический вес» и т.д., воспринимаются на интуитивном уровне даже без знакомства с точными (насколько они тут возможны) определениями. А это позволяет делать подобный анализ актуальных ситуаций предметом широкого обсуждения.

В силу того, что некоторые решения являются неустойчивыми по отношению к своим параметрам, значения таких параметров необходимо определять с высокой точностью. Для этого требуется комплекс методик, основанных не только на обработке больших объемов статистических данных, но и на разносторонних социологических исследованиях.

В настоящее время реалистичной выглядит постановка задачи, состоящая в использовании математических моделей для прогнозирования возможных сценариев динамики социальных процессов на качественном уровне. В такой формулировке моделирование динамики занимает как бы промежуточный уровень между тем, что изложено здесь, и точным прогнозированием. И все же потребуется выбор значений параметров, которые бы в некотором разумном приближении соответствовали изучаемой ситуации, причем в большинстве случаев продуктивным оказывается использование относительных величин.

Так, конечно, не получить достоверных данных о будущем развитии событий, но, скорее всего, можно составить более или менее адекватную картину того, что и как может произойти. А это уже не мало.

При этом для достижения успеха отдельные информационные воздействия необходимо рассматривать как части единой информационной операции, точно так же, как артобстрел или авиационные атаки можно рассматривать как согласованные части военной операции.

Отметим, что информационным операциям присущи такие основные особенности:

- информационные операции – это междисциплинарный набор методов и технологий в таких областях, как информатика, социология, психология, международные отношения, коммуникации, военная наука;
- до сих пор не существует стандартов проведения информационных операций;
- в развитии технологий информационных операций заинтересованы не только оборонные ведомства, но и многие правительственные и коммерческие организации;
- задача формирования научного подхода к информационным операциям является насущной и актуальной.

При проведении информационных операций существенно выявление содержания (знаний), вкладываемого в информацию, с учетом самых разнообразных аспектов – социальных, политических, религиозных, исторических, экономических, психологических, ментальных, культурных, присущих различным слоям общества. Поэтому в настоящее время имеет смысл рассматривать информационные операции шире, как операции, базирующиеся на знаниях

(Knowledge Operations) [Burke, 2001].

Обычная сетевая информационная атака в веб-среде сегодня производится следующим образом: как правило, создается и некоторое время функционирует веб-сайт (назовем его «первоисточником»), при этом он публикует вполне корректную информацию. В час «Х» на его странице появляется документ, обычно компромат на объект атаки, достоверный либо сфальсифицированный. Затем происходит так называемая «отмывка информации». Документ перепечатывают интернет-издания двух типов – заинтересованные в атаке и те, кому попросту не хватает информации для заполнения своего информационного поля. В случае претензий все перепечатывающие издания ссылаются на «первоисточник» и, в крайнем случае, по просьбе/требованию объекта атаки удаляют со своих веб-сайтов информацию. Первоисточник при необходимости также снимает информацию либо вовсе ликвидируется (после чего оказывается, что он зарегистрирован в Интернет на несуществующее лицо). Вместе с тем информация уже разошлась, задача первоисточника выполнена, атака стартовала.

Современное информационное пространство представляет собой уникальную возможность получения любой информации по выбранному вопросу при условии наличия соответствующего инструментария, применение которого позволяет анализировать взаимосвязь возможных событий или событий, которые уже происходят, с информационной активностью определенного круга источников информации. С другой стороны, при ретроспективном анализе любого процесса или явления интерес представляют определенные характеристики его развития, а именно:

- количественная динамика, присущая процессу или явлению, например, количество событий в единицу времени, или количество сообщений, имеющих отношение к нему;

- определение критических, пороговых точек, которые соответствуют количественной динамике явления;
- определение проявлений в критических точках, например, выявление основных сюжетов публикаций в СМИ относительно выбранного процесса или явления;
- после выявления основных проявлений явления в критических точках, эти проявления ранжируются, и исследуется динамика развития отдельных определенных проявлений до и после определенных критических точек;
- осуществляется статистический, корреляционный и фрактальный анализ общей динамики и динамики отдельных проявлений, на основе которых осуществляются попытки прогнозирования развития явления и отдельных его проявлений.

Для исследования взаимосвязи реальных событий и публикаций о них в сети Интернет авторами использовалась система InfoStream, обеспечивающая интеграцию и мониторинг сетевых информационных ресурсов.

Количество веб-публикаций в день по какой-либо теме, в особенности изменения (динамика) этой величины порой позволяют даже непрофессионалам в предметной области делать более-менее точные выводы.

Получить данные подобной динамики можно, например, ежедневно заходя на сайты интеграторов новостей (news.yandex.ru, webground.su, uaport.net). Конечно, в лучшем положении пользователи профессиональных систем мониторинга типа Интегрум или InfoStream. Именно на основе последней системы получена удивительная статистика по количеству веб-публикаций по тематике эпидемий гриппа в разные периоды.

В качестве примера рассмотрим информационную кампанию, направленную против «Проминвестбанка», которая началась в конце сентября 2008 г.

С помощью системы контент-мониторинга InfoStream (www.infostream.ua) [Григорьев, 2007], сканирующей все основные информационные веб-сайты Украины в режиме реального времени, была определена динамика публикаций на веб-сайтах сообщений, в которых упоминался «Проминвестбанк» за три месяца – сентябрь, октябрь и ноябрь (рис. 73). Эта динамика свидетельствует о небольшом количестве публикаций за первую половину сентября, однако затем пошел ряд публикаций, компрометирующих председателя правления В. Матвиенко, что вызвало относительно небольшой резонанс.

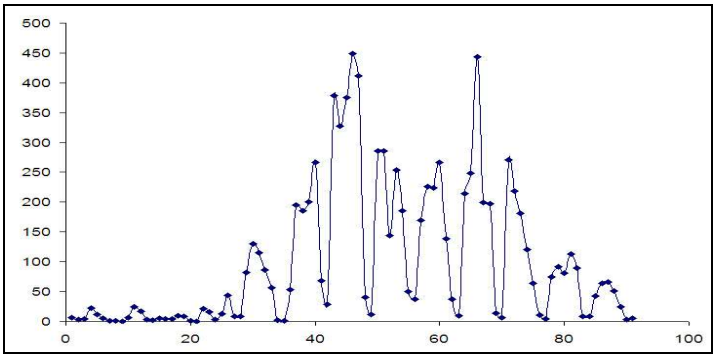
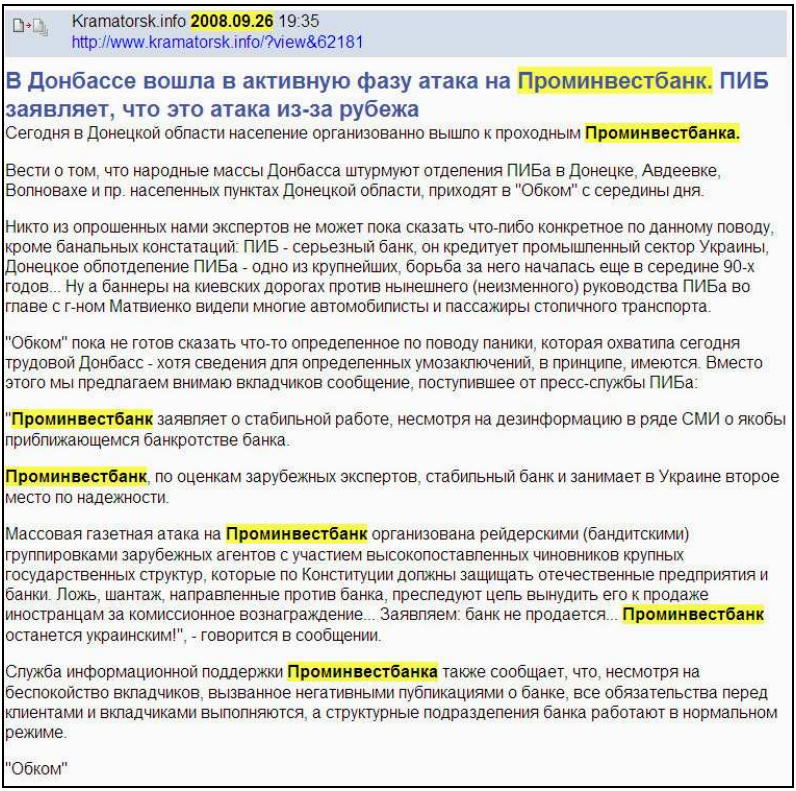


Рис. 73 – Динамика публикаций по теме «Проминвестбанк» за три месяца 2008 г.

Как оказалось впоследствии, эти публикации были лишь «артподготовкой». 26 сентября появились первые сообщения о возможном банкротстве банка (рис. 74), количество которых вполне соответствовало лавинообразному процессу, ограниченному лишь числом веб-сайтов, способных публиковать подобную информацию. Впрочем, этот процесс вышел на стабильно-средний уровень к декабрю 2008 г.

Нельзя утверждать, что лишь информационная атака через сеть Интернет привела Проминвестбанк к

печальному состоянию, однако именно первые тревожные сообщения подорвали доверие многих вкладчиков, заставили их массово забирать свои сбережения из банка.



Kramatorsk.info **2008.09.26** 19:35
http://www.kramatorsk.info/?view&62181

В Донбассе вошла в активную фазу атака на Проминвестбанк. ПИБ заявляет, что это атака из-за рубежа

Сегодня в Донецкой области население организованно вышло к проходным **Проминвестбанка.**

Вести о том, что народные массы Донбасса штурмуют отделения ПИБа в Донецке, Авдеевке, Волновахе и пр. населенных пунктах Донецкой области, приходят в "Обком" с середины дня.

Никто из опрошенных нами экспертов не может пока сказать что-либо конкретное по данному поводу, кроме банальных констатаций: ПИБ - серьезный банк, он кредитует промышленный сектор Украины, Донецкое облотделение ПИБа - одно из крупнейших, борьба за него началась еще в середине 90-х годов... Ну а баннеры на киевских дорогах против нынешнего (неизменного) руководства ПИБа во главе с г-ном Матвиенко видели многие автомобилисты и пассажиры столичного транспорта.

"Обком" пока не готов сказать что-то определенное по поводу паники, которая охватила сегодня трудовой Донбасс - хотя сведения для определенных умозаключений, в принципе, имеются. Вместо этого мы предлагаем внимаю вкладчиков сообщение, поступившее от пресс-службы ПИБа:

"**Проминвестбанк** заявляет о стабильной работе, несмотря на дезинформацию в ряде СМИ о якобы приближающемся банкротстве банка.

Проминвестбанк, по оценкам зарубежных экспертов, стабильный банк и занимает в Украине второе место по надежности.

Массовая газетная атака на **Проминвестбанк** организована рейдерскими (бандитскими) группировками зарубежных агентов с участием высокопоставленных чиновников крупных государственных структур, которые по Конституции должны защищать отечественные предприятия и банки. Ложь, шантаж, направленные против банка, преследуют цель вынудить его к продаже иностранцам за комиссионное вознаграждение... Заявляем: банк не продается... **Проминвестбанк** останется украинским!", - говорится в сообщении.

Служба информационной поддержки **Проминвестбанка** также сообщает, что, несмотря на беспокойство вкладчиков, вызванное негативными публикациями о банке, все обязательства перед клиентами и вкладчиками выполняются, а структурные подразделения банка работают в нормальном режиме.

"Обком"

Рис. 74 – Одно из первых тревожных сообщений

Через три дня 30 сентября 2008 г. появилось сообщение, что для спасения Проминвестбанка Национальный Банк Украины (НБУ) решил выделить ему 5 млрд. гривен рефинансирования, а 5 декабря появилось сообщение, что у «Проминвестбанка» появился новый владелец (рис. 75). После этого объемы публикаций о Проминвестбанке существенно сократились, что свидетельствует не столько об его оздоровлении, сколько о системном кризисе банковской

системы Украины, «уронившему» многие другие кредитные и банковские учреждения.

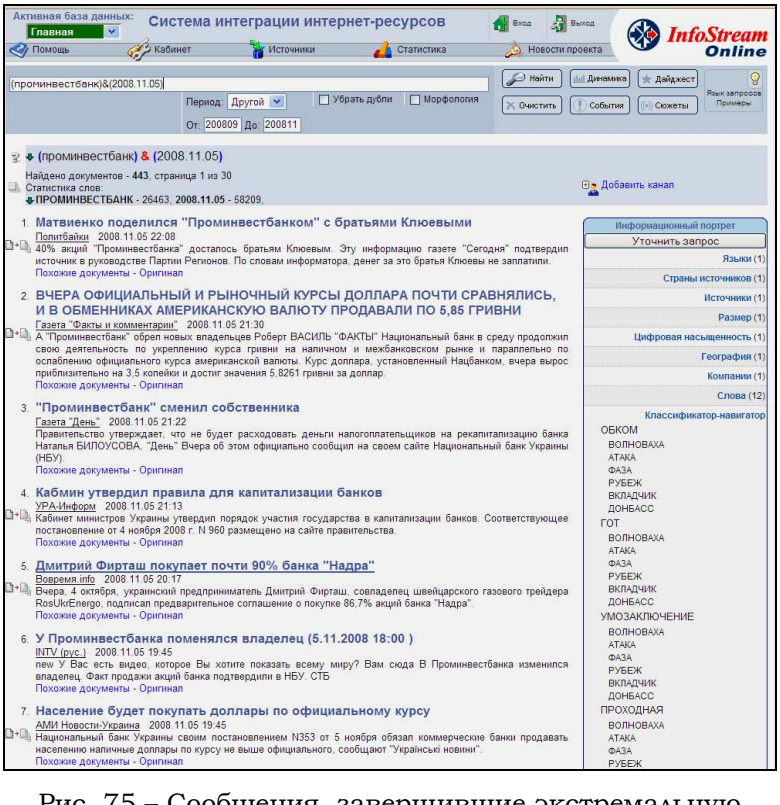


Рис. 75 – Сообщения, завершившие экстремальную динамику интенсивности публикаций по теме «Проминвестбанк»

Буквально через неделю после описанных выше событий в Украине произошла еще одна публичная знаковая информационная атака, в этот раз на рынке страхования. Это была настоящая информационная операция против Национальной акционерной страховой компании (НАСК) «Оранта». В этом случае первоисточником компромата оказался не веб-сайт, а информационное сообщение, разосланное электронной почтой тысячам пользователей Интернет. В результате применения специальных технических приемов, оно

разошлось с обозначением адреса пресс-службы объекта атаки. Итак, 10 декабря 2008 года в районе 11:30 в виде спама было разослано информационное сообщение, в котором говорилось о том, что страховая компания «Оранта» заявляет о банкротстве. По предварительным данным, информация разлетелась по 1000 адресам, естественно, данные попали к конкурентам и в СМИ. В сообщении говорилось, что компания с 31 декабря 2008 года прекращает выполнять взятые перед клиентами обязательства.

В связи со случившимся НАСК «Оранта» обратилась в правоохранительные органы с просьбой расследовать данный инцидент и наказать виновных. Произошедшее с НАСК «Оранта» очень напоминало ситуацию с Проминвестбанком, с этим согласились многочисленные эксперты. Ведь как банковский бизнес, так и страховой основываются на доверии клиентов, которое легче всего подрывается именно информационными атаками. По словам Олега Спилки, председателя наблюдательного совета НАСК «Оранта», «Это мероприятие готовилось целенаправленно для того, чтобы дискредитировать страховую компанию и подорвать ее репутацию». Не вдаваясь в детали возможных целей атаки (смена владельцев, борьба за блокирующий пакет акций, уничтожение компании и т.п.), с помощью ретроспективного анализа проследим за динамикой публикаций в сети Интернет, в которых упоминалась НАСК «Оранта».

На рис. 76 приведена посуточная динамика количества соответствующих публикаций. На этой диаграмме, кроме всего прочего, отчетливо виден спад интенсивности публикаций по данной теме в начале декабря 2008 г., что вполне можно воспринимать как некоторое «затишье перед бурей».

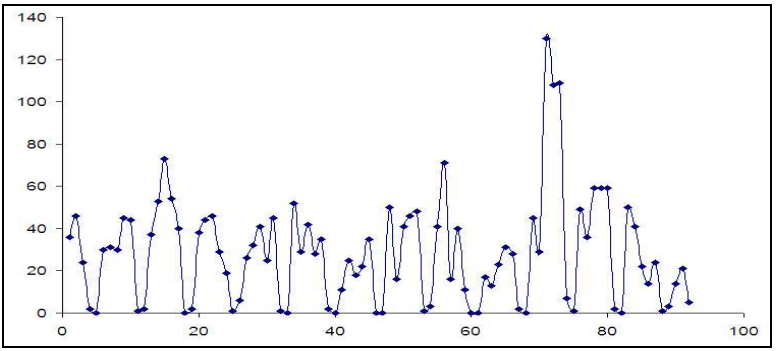


Рис. 76 – Интенсивность публикаций в Интернет по теме «Оранта»

Для анализа временных рядов в рамках исследования применялся ΔL-метод. На рис. 77 представлена скейлограмма динамики рассматриваемого процесса с помощью метода (ΔL-метода) за второе полугодие 2008 года. Несмотря на отдельные пики в 16 и 55 день квартала, все же наибольший интерес представляет экстремум, приходящийся именно на 10–12 декабря.

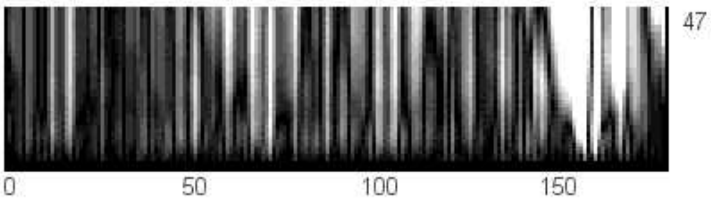
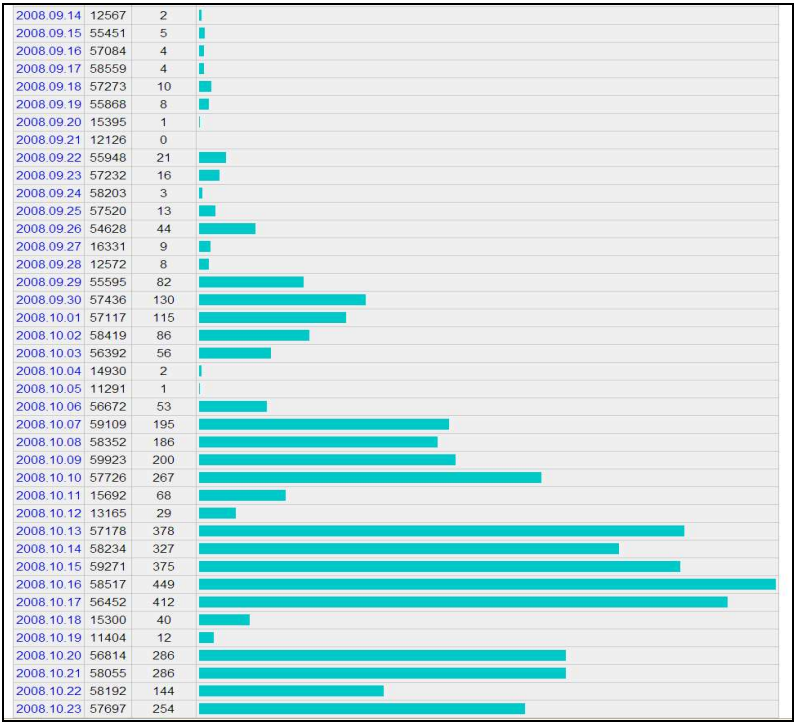


Рис. 77 – ΔL-диаграмма ряда публикаций по теме «Оранта»

Более детальная статистика публикаций по теме «Оранта» за декабрь 2008 года получена через интерфейс пользователя системы контент-мониторинга InfoStream (рис. 78).



2008.09.14	12567	2
2008.09.15	55451	5
2008.09.16	57084	4
2008.09.17	58559	4
2008.09.18	57273	10
2008.09.19	55868	8
2008.09.20	15395	1
2008.09.21	12126	0
2008.09.22	55948	21
2008.09.23	57232	16
2008.09.24	58203	3
2008.09.25	57520	13
2008.09.26	54628	44
2008.09.27	16331	9
2008.09.28	12572	8
2008.09.29	55595	82
2008.09.30	57436	130
2008.10.01	57117	115
2008.10.02	58419	86
2008.10.03	56392	56
2008.10.04	14930	2
2008.10.05	11291	1
2008.10.06	56672	53
2008.10.07	59109	195
2008.10.08	58352	186
2008.10.09	59923	200
2008.10.10	57726	267
2008.10.11	15692	68
2008.10.12	13165	29
2008.10.13	57178	378
2008.10.14	58234	327
2008.10.15	59271	375
2008.10.16	58517	449
2008.10.17	56452	412
2008.10.18	15300	40
2008.10.19	11404	12
2008.10.20	56814	286
2008.10.21	58055	286
2008.10.22	58192	144
2008.10.23	57697	254

Рис. 78 – Детальная диаграмма интенсивности публикаций по теме «Оранта»

Проследим за ходом информационной операции, рассматривая сообщения, публикуемые в разные промежутки времени.

На рис. 79 приведен список публикаций по теме «Оранта» в течение первых часов атаки. По словам Олега Спилки, в течение двух часов с начала атаки все почтовые серверы НАСК «Оранта» были выведены из строя, поэтому опровержение в сети задержалось.

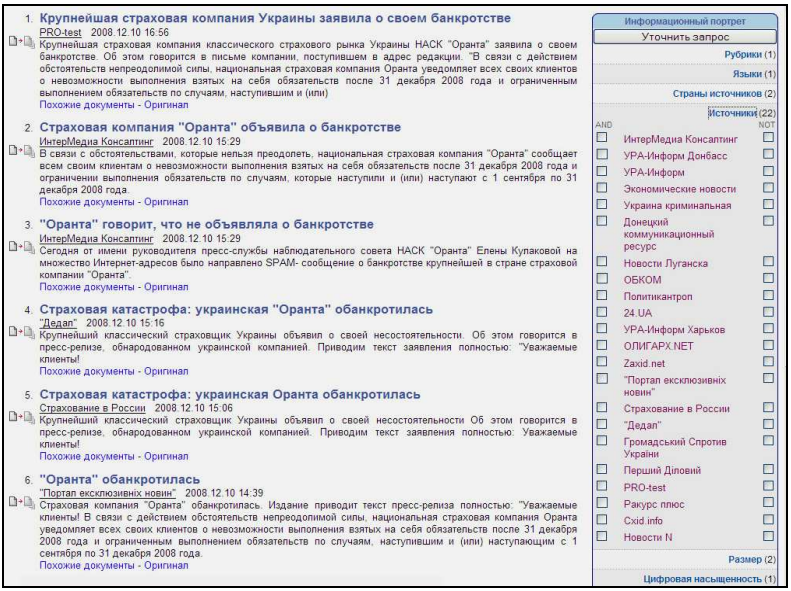
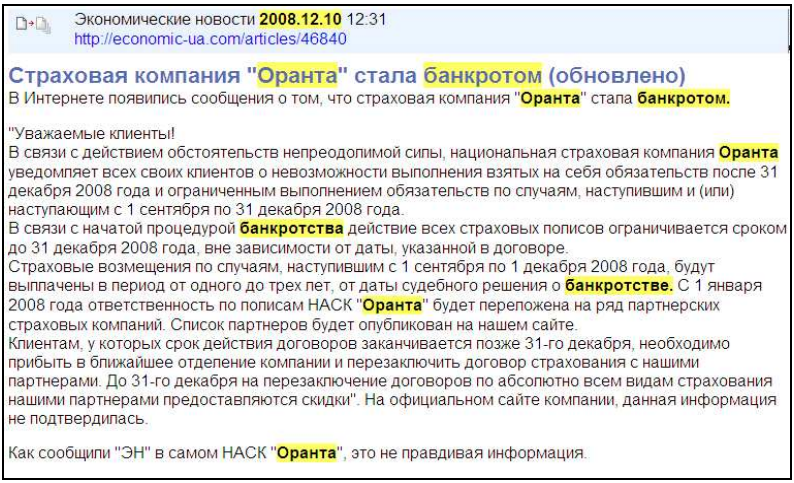


Рис. 79 – Самые «оперативные» источники

В 12:31 на сайте «Экономические новости» появляется странное «обновленное» сообщение с парадоксальным последним предложением (рис. 80).

На рис. 81 приведен список публикаций, посвященных опровержению сообщения о банкротстве НАСК «Оранта» за следующий день (11 декабря), а также наиболее активных источников, опубликовавших эти сообщения. Безусловный интерес аналитиков вызывает сравнение источников, приведенных на рис. 79 и 81.

Далее руководство НАСК «Оранта» опубликовало в Интернете первые опровержения, не спеша обвинять конкурентов в происшедшем, а затем все же признав атаку целенаправленной и выгодной третьим лицам.



Экономические новости **2008.12.10** 12:31
http://economic-ua.com/articles/46840

Страховая компания "Оранта" стала банкротом (обновлено)

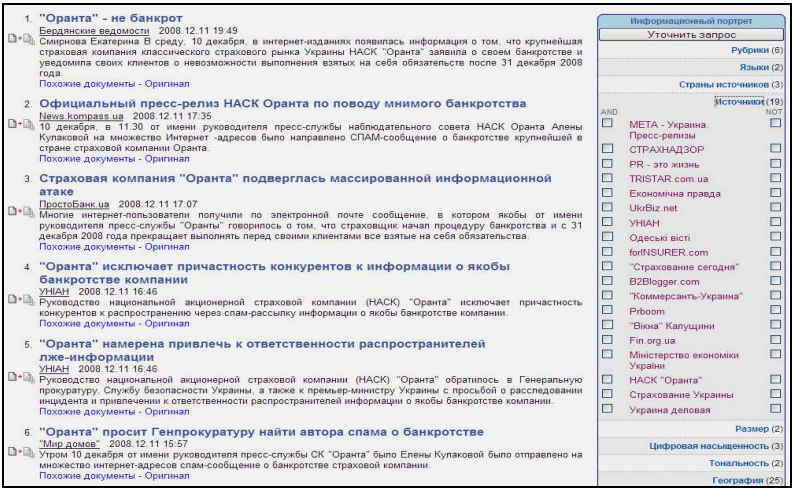
В Интернете появились сообщения о том, что страховая компания "**Оранта**" стала **банкротом.**

"Уважаемые клиенты!
В связи с действием обстоятельств непреодолимой силы, национальная страховая компания **Оранта** уведомляет всех своих клиентов о невозможности выполнения взятых на себя обязательств после 31 декабря 2008 года и ограниченным выполнением обязательств по случаям, наступившим и (или) наступающим с 1 сентября по 31 декабря 2008 года.
В связи с начатой процедурой **банкротства** действие всех страховых пописов ограничивается сроком до 31 декабря 2008 года, вне зависимости от даты, указанной в договоре.
Страховые возмещения по случаям, наступившим с 1 сентября по 1 декабря 2008 года, будут выплачены в период от одного до трех лет, от даты судебного решения о **банкротстве.** С 1 января 2008 года ответственность по полисам НАСК "**Оранта**" будет перепожена на ряд партнерских страховых компаний. Список партнеров будет опубликован на нашем сайте.
Клиентам, у которых срок действия договоров заканчивается позже 31-го декабря, необходимо прибыть в ближайшее отделение компании и перезаключить договор страхования с нашими партнерами. До 31-го декабря на перезаключение договоров по абсолютно всем видам страхования нашими партнерами предоставляются скидки". На официальном сайте компании, данная информация не подтвердилась.

Как сообщили "ЭН" в самом НАСК "**Оранта**", это не правдивая информация.

Рис. 80 – Опровержение?

Дальнейший спад публикаций по теме НАСК «Оранта» и возвращение их на нормальный «средний» уровень свидетельствует о том, что компания своими осторожными и точными действиями смогла с успехом противостоять информационной операции.



1. **"Оранта" - не банкрот**
Бердянские ведомости 2008.12.11 19:49
Смирнова Екатерина В среду, 10 декабря, в интернет-изданиях появилась информация о том, что крупнейшая страховая компания классического страхового рынка Украины НАСК "Оранта" заявила о своем банкротстве и уведомила своих клиентов о невозможности выполнения взятых на себя обязательств после 31 декабря 2008 года.
Похожие документы - Оригинал

2. **Официальный пресс-релиз НАСК Оранта по поводу мнимого банкротства**
News.kompass.ua 2008.12.11 17:35
10 декабря, в 11.30 от имени руководителя пресс-службы наблюдательного совета НАСК Оранта Алены Кулаковой на множество Интернет -адресов было направлено СПАМ-сообщение о банкротстве крупнейшей в стране страховой компании Оранта.
Похожие документы - Оригинал

3. **Страховая компания "Оранта" подверглась массированной информационной атаке**
ПростоБанк.ua 2008.12.11 17:07
Многие интернет-пользователи получили по электронной почте сообщение, в котором якобы от имени руководителя пресс-службы "Оранты" говорилось о том, что страховщик начал процедуру банкротства и с 31 декабря 2008 года прекращает выполнять перед своими клиентами все взятые на себя обязательства.
Похожие документы - Оригинал

4. **"Оранта" исключает причастность конкурентов к информации о якобы банкротстве компании**
УНІАН 2008.12.11 16:46
Руководство национальной акционерной страховой компании (НАСК) "Оранта" исключает причастность конкурентов к распространению через спам-рассылку информации о якобы банкротстве компании.
Похожие документы - Оригинал

5. **"Оранта" намерена привлечь к ответственности распространителей лжеинформации**
УНІАН 2008.12.11 16:46
Руководство национальной акционерной страховой компании (НАСК) "Оранта" обратилось в Генеральную прокуратуру, Службу безопасности Украины, а также к премьер-министру Украины с просьбой о расследовании инцидента и привлечении к ответственности распространителей информации о якобы банкротстве компании.
Похожие документы - Оригинал

6. **"Оранта" просит Генпрокуратуру найти автора спама о банкротстве**
"Мир домов" 2008.12.11 15:57
Утром 10 декабря от имени руководителя пресс-службы СК "Оранта" было Елены Кулаковой было отправлено на множество интернет-адресов спам-сообщение о банкротстве страховой компании.
Похожие документы - Оригинал

Информационный портрет
Уточнить запрос
Рубрики (6)
Языки (2)
Страны источников (3)
Источники (19)
AND NOT
META - Украина. Пресс-релизы
СТРАХНАДЗОР
PR - это жизнь
TRISTAR.com.ua
Економічна правда
UkrBiz.net
УНІАН
Одеські вісті
forINSURER.com
"Страхование сегодня"
B2Blogger.com
"Коммерсантъ-Украина"
Prboom
"Вікна" Калущини
Fin.org.ua
Міністерство економіки України
НАСК "Оранта"
Страхование Украины
Украина деловая
Размер (2)
Цифровая насыщенность (3)
Тональность (2)
География (25)

Рис. 81. Сообщения с опровержением

5.4. Сетевая мобилизация

Социальные сети сегодня вызывают большой интерес, в частности потому, что в процессе развития они приобретают качественно новые свойства, среди которых следует выделить способность к проведению сетевой мобилизации. Сетевая мобилизация обычно рассматривается как средство объединения усилий участников социальных сетей для решения некоторых проблем, например, организации массовых выступлений, отражения агрессии, помощи пострадавшим и т.п.

Вопросы сетевой мобилизации были актуальными всегда. Такие социальные сети, как партизанские подполья, кружки революционеров, заговорщиков, порой именно путем сетевой мобилизации добивались своих целей.

Сегодня с вопросами сетевой мобилизации ассоциируются преимущественно социальные сети в Интернете, которые все чаще выступают как средства информационного управления и манипулирования людьми, обществом.

Возможность влияния в социальной сети зависит от репутации [Губанов, 2009] тех, кто его оказывает. Репутация рассматривается как некоторая весовая величина, которая растет, если выбор объекта влияния (участника) совпадает с тем, что от него ожидается другими, или снижается при неэффективном управлении.

Возможности сетевой мобилизации зависят от:

1) структуры сети, ее топологии, параметров, динамики информации, циркулирующей в ней;

2) возможности восстановления связей в сети после деструктивного воздействия на них, а также учета скрытых (латентных) связей, не включенных в заданную изначально в явном виде топологию сети;

3) возможности и вероятности восприятия информации узлами сети;

4) возможности преобразования/переработки информации в узлах сети.

Вопросы структуры и топологии социальных сетей

Возможности сетевой мобилизации напрямую связаны с такими свойствами сетей, как связность, кластеризация, средний кратчайший путь между вершинами и т.п. Известно, что социальные сети могут быть представлены в виде динамических сетей.

В случае сетевой мобилизации необходимо учитывать ряд особенностей [Stohl, 2003]:

- ребра сети не являются статичными, они могут развиваться на нескольких уровнях, в том числе скрытых, латентных;
- реальные сети не являются четко иерархическими;
- реальные сети могут разъединяться или объединяться, при этом отдельные подсети могут функционировать как полнофункциональные.

При моделировании сетевой мобилизации возникает необходимость учета факторов, имеющих место в реальных социальных сетях [Губанов, 2009]:

- наличие собственного мнения агентов, которое может изменяться под влиянием мнений других;
- целенаправленное поведение агентов;
- различная репутация участников сети (агентов) – различная значимость их мнений;
- различная степень подверженности агентов влиянию;

- различный порог чувствительности к изменению мнения окружающих;
- наличие внешнего воздействия;
- асимметричная информированность агентов и т.п.

Для социальных сетей выявлен ряд эффектов, которые имеют важное значение при реализации сетевой мобилизации, остановимся на некоторых из них.

«Малые миры». В 1967 г. С. Милгран в результате масштабных экспериментов определил, что существует цепочка знакомств, в среднем длиной шесть, между любыми двумя гражданами США [Milgram, 1967]. Сеть знакомств получила название «малого мира». Сетевые структуры, соответствующие свойствам малых миров имеют следующие типичные свойства: малая средняя длина пути и большая кластеризация (что присуще сетям с регулярной структурой).

Если в сеть, имеющую структуру малого мира «вбросить» мобилизующие идеи, то они будут распространяться там, как эпидемия. При точном выборе подходящих идей возникает мобилизация как массовая социальная реакция.

«Слабые связи». Существует класс социальных сетей, обладающих так называемыми «слабыми» связями, например, сети отношений с дальними знакомыми и коллегами. «Слабые» социальные связи оказываются более важными для существования социальной сети, чем связи «сильные». Оказалось, что именно слабые связи является тем феноменом, который связывает сеть в единое целое. Если же слабые связи проигнорировать, то сеть распадется на отдельные фрагменты.

«Клуб богатых». Во многих социальных сетях наблюдается такая тенденция, как хорошая связность между узлами-концентраторами. Это явление, известное под названием элитарность (или феномен «клуб богатых»

– rich-club phenomenon), может быть охарактеризовано коэффициентом элитарности [Zhou, 2004]. Анализ топологии веб, в частности, показал, что узлы с большой степенью выходных гиперссылок имеют больше связей между собой, чем с узлами с малой степенью.

«Структура сообщества». Социальные сети характеризуются наличием так называемой «структуры сообщества», т.е. существуют группы узлов-агентов, которые имеют высокую плотность ребер между собой, при том, что плотность ребер между отдельными группами – низкая.

«Клеточные сети». Социальные сети часто характеризуются как клеточные [Frantz, 2005] – созданные из почти независимых клеток. Клеточные сети имеют такие свойства, как избыточность, наличие тесно связанных клеток (4-6 вершин), отсутствие управления вертикальным способом (нечеткие директивы), отсутствие планирования (формирование за счет локальных ограничений), возможность эволюционирования в ответ на деструктивную деятельность [Sageman, 2004].

«Ценность сети». С точки зрения возможности мобилизации в сети применяют понятие ценности сети [Бреер, 2009] как: «это потенциальная доступность участников сети (узлов, агентов), с которыми любой может связаться в случае необходимости». Д. Сарнов определил, что ценность сетей общественного вещания растет пропорционально количеству слушателей n. Р. Меткалф определил, что ценность социальной сети растет как $n\,(n-1)$ так как каждый агент социальной сети может быть связан с $n-1$ другими агентами. Д. Род прибавил к выражению ценности социальной сети еще одну составляющую, связанную с объединением агентов сети в группы. Эта составляющая составляет $2^n - n - 1$ и определяется как количество подмножеств множества из n агентов за исключением единичных элементов и пустого множества. Известны оценки ценности сети как $n \log_2(n)$.

Для ценности социальной сети предлагается

описание, которое отражает свойство аддитивности: ценность объединения двух сетей должна быть равна сумме ценностей этих сетей, т.е. для функции ценности должна выполняться формула:

$$f(m_1 m_2) = f(m_1) + f(m_2),$$

где m_1 и m_2 – количества конфигураций первой и второй сети, соответственно. Если существует только одна конфигурация связей, то будем считать, что ценность такой сети равняется нулю, то есть $f(1) = 0$. Известно, что существует лишь единственная функция, которая удовлетворяет названным требованиям – логарифм. Вместе с тем, если количество возможных конфигураций для сети из n узлов оценивать как 2^n, то $\log_2(2^n) = n$ и мы возвращаемся к результатам Сарнова.

Деструктивные влияния и восстановление латентных связей

Большая безмасштабная сеть допускает случайные удаления до 80% ее узлов и только потом такая сеть распадается. Причина этого заключается в том, что случайные отказы более вероятны для относительно небольших узлов. Вместе с тем, безмасштабные сети очень уязвимы с точки зрения целенаправленных разрушений их концентраторов. Атаки, которые мгновенно уничтожают лишь 5-15% концентраторов подобных сетей, могут разрушить всю сеть.

Безмасштабные сети достаточно подвержены воздействию эпидемий (в случаях социальных сетей в качестве «инфекции» могут рассматриваться идеологические влияния, технические инновации и т.д.). Произвольно эпидемия в сети должна преодолеть некоторый критический порог (количество зараженных узлов) и только тогда она может распространяться на всю сеть. Если заражено количество узлов, меньшее этого порога, то эпидемия угасает. Данные, приведенные в работе [Pastor-Satorras, 2001], показывают, что в безмасштабной сети эпидемический порог практически равен нулю.

Свойства сложных сетей определяют тактику их разрушения, которая предусматривает такие этапы как анализ и планирование, практически одновременная нейтрализация узлов-концентраторов, последовательное уничтожение других узлов в порядке убывания соответствующих им весов.

Зная, например, только часть связей иерархической сети, можно с высокой вероятностью восстановить сведения о недостающих звеньях. Даже не имея полного описания сети, можно получать репрезентативную выборку связей и по ней пытаться достраивать всю сеть. С учетом природы латентных связей сегодня исследуются многочисленные сети, которые порождаются разнообразными объектами (партиями, компаниями, персонами). Это позволяет сетевым экспертам-аналитикам делать выводы относительно общих интересов отдельных групп объектов во времени, выявлять ключевые элементы сетей, пренебрегать несущественными элементами и т.п.

Известно, что матрицы взаимосвязей объектов являются одной из форм представления сетей. При различных подходах к построению матриц взаимосвязей объектов – это, как правило, симметричные матрицы, элементы которых – коэффициенты взаимосвязей.

Рассмотрим одно из формальных определений матрицы взаимосвязей объектов.

Обозначим p_i $(i = 1,...,K)$ – объект, $d^{(j)}$ $(j = 1,...,N)$ – документ, $d^{(j)} \in D$ – массив документов, $e_i^{(j)}$ – признак соответствия объекта p_i документу $d^{(j)}$:

$$e_i^{(j)} = \begin{cases} 1, & p_i \in d^{(j)} \\ 0, & p_i \notin d^{(j)} \end{cases}$$

Можно определить уровень связи объектов p_i и p_k:

$$M_{ik} = \sum_{j=1}^{N} e_i^j e_k^j \ .$$

Если перейти к анализу реальной ситуации, то связь между объектами можно рассматривать как вероятностную. Соответственно нормализовав значения элементов матрицы можно перейти к так называемой «матрице нечетких связей», элементами которых являются вероятности связей между объектами.

Предусматривается, что $p_{i,j}$ – оценка вероятности связи объектов i и j. В общем случае предусматривается, что эта оценка экспертная, независимая от других узлов сети. Эти оценки можно было бы уточнить, учитывая не только прямые связи, но и их связи через третьи, четвертые и т.д. узлы. Допустим, что узлы 1 и 2, связаны непосредственно друг с другом, а также через узел 3 (рис. 82). Соответствующие оценки вероятностей связей составляют $p_{1,2}$ $p_{1,3}$.

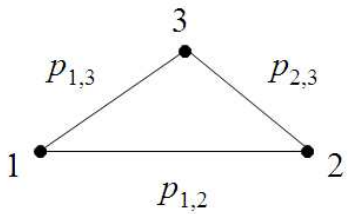


Рис. 82 – Начальные оценки вероятности связей

Тогда можно сделать следующую оценку «нечеткой» вероятности того, что связки между узлами 1 и 2 не существуют:

$$\overline{p}_{1,2}^{(1)} = (1 - p_{1,2})(1 - p_{1,3} p_{3,2}) \ .$$

Соответственно, новая оценка вероятности связи между узлами 1 и 2 составляет:

$$p_{1,2}^{(1)} = 1 - (1 - p_{1,2})(1 - p_{1,3} p_{3,2}) \ .$$

Формула учета всех связей через «третьи узлы» имеет вид:

$$p_{i,j}^{(1)} = 1-(1-p_{i,j})\prod_{k\neq i,j}(1-p_{i,k}p_{k,j}).$$

Очевидно, при $p_{i,j}\in[0,1]$ для любых i и j будет иметь место то, что $p_{i,j}^{(1)}\in[0,1]$ для любых i и j. Действительно, это следует из того, что величина $(1-p_{i,j})\prod_{k\neq i,j}(1-p_{i,k}p_{k,j})$ не больше единицы и не меньше нуля как произведение неотрицательных сомножителей, каждый из которых меньше единицы.

При этом всегда $p_{i,j}^{(1)}\geq p_{i,j}$. Докажем это, для чего введем обозначение $\alpha=\prod_{k\neq i,j}(1-p_{i,k}p_{k,j})$. Очевидно $\alpha\in[0,1]$. Необходимо доказать, что $p_{i,j}^{(1)}-p_{i,j}\geq 0$. Это следует из выражения:

$$p_{i,j}^{(1)}-p_{i,j}=1-\alpha(1-p_{i,j})-p_{i,j}=(1-p_{i,j})(1-\alpha).$$

Каждый из полученных сомножителей не негативный, следовательно, их произведение также не негативно. Можно также оценить вероятность с учетом связей через 4-й, 5-й и так далее узлы, модифицировав функцию расчета $p_{i,j}^{(1)}$ таким образом:

$$p_{i,j}^{(1)} = 1-(1-p_{i,j})\prod_{k\neq i,j}(1-p_{i,k}p_{k,j})\times$$

$$\times\prod_{k\neq i\neq l\neq j}(1-p_{i,k}p_{k,l}p_{l,j})\prod_{k\neq i\neq l\neq m\neq j}(1-p_{i,k}p_{k,l}p_{l,m}p_{m,j})\ldots$$

Полученная матрица будет отражать не только явные связи, выраженные оценками вероятности, но и связи 2-го, 3-го и т.д. уровней. На практике

сомножители, начиная уже с $\prod_{k\neq i\neq l\neq j}(1-p_{i,k}\,p_{k,l}\,p_{l,j})$ оказываются настолько близкими к единице, что их обычно можно не учитывать в практических расчетах.

К полученной в результате матрице, элементами которой являются $p_{i,j}^{(1)}$, также можно применить приведенный выше алгоритм, рассматривая всего лишь как первую итерацию:

$$p_{i,j}^{(m+1)} = 1-(1-p_{i,j}^{(m)})\prod_{k\neq i,j}(1-p_{i,k}^{(m)}\,p_{k,j}^{(m)})\,.$$

Здесь $p_{i,j}^{(m)}$ для любых i и j является монотоно неубывающей функцией от m, кроме того, при достаточно больших n все элементы матрицы $\left\|p_{i,j}^{(m)}\right\|$ кроме диагональных, оказываются близкими к единице. При необходимости проведения нескольких итераций осуществляется нормирование величин $p_{i,j}^{(m)}$ путем возведения их в некоторую степень $\gamma>1$.

Для оценки действенности предложенного подхода строилась матрица, соответствующая сети со степенным распределением веса узлов, из которой удалялись (обнулялись весовые значения) ребра со значениями в среднем диапазоне. При единичном удалении ребер, они практически всегда восстанавливались с помощью приведенного метода за 1 шаг итерации. При удалении 20% ребер они восстанавливались примерно в 75% случаев.

Модель распространения идей, зависящая от вероятности их восприятия

Авторами построена модель распространения мобилизационных идей в сети, построенная на основе концепции клеточных автоматов. Клеточный автомат представляет собой дискретную динамическую систему, совокупность одинаковых клеток, одинаковым образом соединенных между собой. Все клетки образуют сеть

клеточных автоматов. Состояние каждой клетки определяются состоянием клеток, входящих в ее локальную окрестность. Окрестностью конечного автомата с номером j называется множество его «ближайших соседей». Состояние j-го клеточного автомата в момент времени $t+1$, определяется следующим образом:

$$y_j(t+1) = F\left(y_j(t), O(j) < t\right),$$

где F – некоторое правило, $O(j)$ – окрестность, t – такт.

Динамике распространения информации присущи некоторые свойства (в частности, старения), которые и были учтены в модели. Предполагается, что клетка может быть в одном из пяти состояний: 0 – идея не дошла до клетки (клетка окрашивается в белый цвет); 1 – «свежая идея» (клетка окрашивается в черный цвет); 2 – 4 устаревшие сведения (клетки, окрашенные в оттенки серого). Правила распространения идей следующие:

- изначально все поле состоит из белых клеток за исключением одной – черной, которая первой «приняла» идею;
- белая клетка может перекрашиваться только в черный цвет или оставаться белой (она может принимать идею или оставаться «в неведении»);
- белая клетка перекрашивается с некоторой наперед заданной вероятностью p (важнейший параметр модели), если в ее окрестности есть хотя бы одна черная клетка;
- если клетка черная или серая, то она перекрашивается в более светлый оттенок серого или белый цвет (происходит забывание идеи).

Описанная система клеточных автоматов вполне

реалистично отражает процесс распространения идей среди участников мобилизационной сети. Пример работы модели на поле размером 40 х 40 (размеры были выбраны авторами исключительно с целью наглядности) приведен на рис. 83.

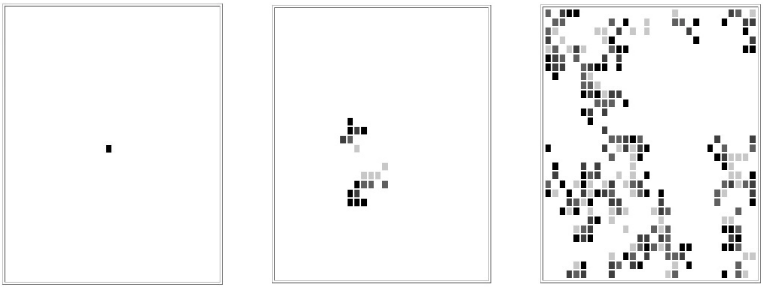


Рис. 83. Процесс эволюции системы клеточных автоматов

Многочисленные эксперименты с данным клеточным автоматом показывают типичные зависимости количества черных клеток от шага эволюции и вероятности (*p*) принятия идеи отдельными клетками (рис. 84).

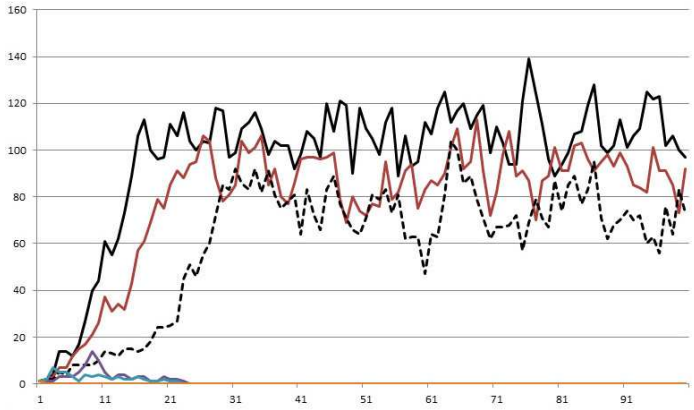


Рис. 84. Количество черных клеток в зависимости от шага эволюции и вероятности принятия идей (различные кривые)

Наиболее быстрый рост и последующие колебания вокруг некоторых уровней насыщения наблюдаются при высокой вероятности принятия идей 30%, 35% и 40% и более. При меньших вероятностях 25% , 20% и 15% идеи в данной модели сети не становятся мобилизационными и быстро забываются.

Возможности преобразования/переработки информации учасниками сети при взаимодействии с другими агентами являются определяющими для обеспечения живучести [Додонов, 2011] мобилизационной идеи, которая, в свою очередь, во многом определяет эффективность (а иногда и возможность) всей сетевой мобилизации.

Социальные сети в настоящее время все больше рассматриваются как онлайновые социальные сети в Интернете, такие как Twitter, Facebook, «Вконтакте» и др. Почему такое большое внимание при изложении уделяется Интернету? Очевидно, большинство человечества не имеет к нему постоянного нецензурированного доступа. Однако ситуация может измениться в корне. Известно, что нововведения внедряются со все большим ускорением. При этом, если социальные сети позволяют осуществлять информационное управление (манипулирование, скрытое управление), то неизбежно возникает и (двойственная) задача – анализ и обеспечение информационной безопасности таких сетей.

5.5. Выявление информационных операций

Для оперативного анализа информационной обстановки с целью выявления информационных операций применяются специализированные системы мониторинга информационного пространства (контент-мониторинга). Такие системы обеспечивают, во-первых, оперативность, которую не могут обеспечить традиционные поисковые системы (время индексации сетевого контента даже лучшими из них составляет от нескольких суток до нескольких недель). Во-вторых,

полноту (как в плане источников, так и представления материалов источников), которую не всегда обеспечивают обычные агрегаторы новостей. И, в-третьих, необходимые аналитические средства, которые позволяют пользователю создавать аналитические отчеты, базирующиеся на публикациях по заданной тематике в необходимый период времени.

В плане профилактики информационных операций следует внимательно следить за динамикой публикаций о целевой кампании, если есть возможность, с учетом тональности этих публикаций, пользоваться доступными аналитическими средствами, например, вейвлет-анализом. При этом следует ориентироваться на возможные модели информационных атак, например, если эта модель охватывает фазы: «фоновые публикации» — «затишье» — «артподготовка» — «затишье» — «атака» (рис. 85), то уже по первым трем компонентам можно с большой вероятностью предсказать грядущие события.

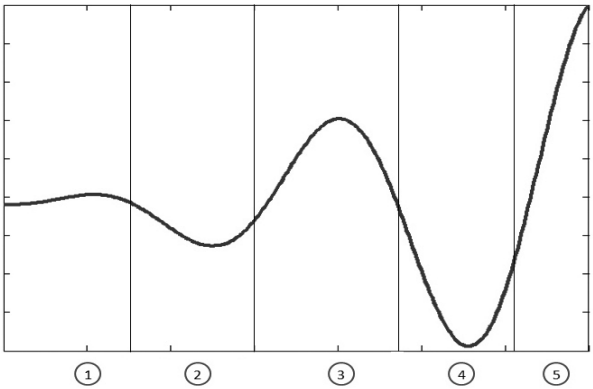


Рис. 85 – Динамика количества тематических сообщений во время проведения информационной операции: 1 – фон; 2 – затишье; 3 – «артподготовка»; 4 – затишье; 5 – атака/триггер роста

Приведенный выше план, очевидно, является идеальным, ориентированным исключительно на данные контент-мониторинга веб-ресурсов.

Конечно, в лучшем положении находятся пользователи профессиональных систем контент-мониторинга. Многие современные информационно-аналитические системы содержат в своем составе средства отображения статистики вхождения в базы данных понятий, соответствующих пользовательским запросам. В частности, авторами использовалась подсистема статистики в рамках системы контент-мониторинга веб-пространства InfoStream, реализующая данную функциональность.

При изучении трендов информационных операций в качестве временных рядов рассматриваются именно ряды по количеству тематических публикаций за определенный промежуток времени (чаще всего – за сутки), соответствующие этим информационным операциям. Поэтому для выявления трендов исследуются информационные потоки, соответствующие тематикам информационных операций – тематические информационные потоки.

Приведенные в [Горбулін, 2009] тренды сообщений, соответствующие этапам информационной операции, представлены на рис. 86. При этом аналитики уже по первым трем компонентам (фон» – «затишье» – «артподготовка») могут с большой вероятностью предсказать будущие события.

Следует отметить, что подобная динамика количества тематических сообщений при проведении информационных операций хорошо описывается известным уравнением распространения электромагнитных волн $y = A + Bx\sin(x)$, где x – время, A и B – константы, определяемые эмпирически.

Как известно, в настоящее время инновационная деятельность также косвенно измеряется количеством публикаций, относящимся к инновациям, существует несколько моделей инновационных процессов, среди которых можно выделить модель диффузии инноваций [Bhargava, 1993]. Вместе с тем, внедрение инноваций также можно считать информационными операциями.

Поэтому обратимся к результатам соответствующих исследований. На рис. 53 приведена обоснованная в [Хорошевский, 2012] диаграмма количества публикаций, соответствующая тренду инновационной деятельности.

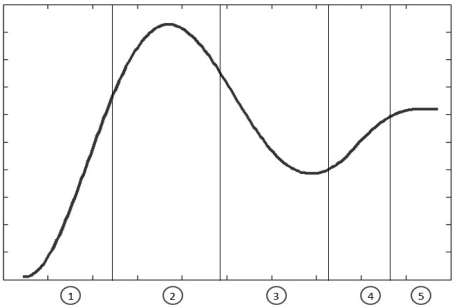


Рис. 86 – Диаграмма количества публикаций, соответствующих тренду инновационной деятельности: 1 – атака/триггер роста; 2 – пик завышенных ожиданий; 3 – утрата иллюзий; 4 – общественное осознание; 5 – продуктивность/фон

Следует отметить, что подобная динамика количества тематических сообщений при проведении информационных операций хорошо описывается известным уравнением распространения электромагнитных волн $y = A + Bx\sin(x)$, где x – время, A и B – константы, определяемые эмпирически.

Как известно, в настоящее время инновационная деятельность также косвенно измеряется количеством публикаций, относящимся к инновациям, существует несколько моделей инновационных процессов, среди которых можно выделить модель диффузии инноваций [Bhargava, 1993]. Вместе с тем, внедрение инноваций также можно считать информационными операциями. Поэтому обратимся к результатам соответствующих исследований. На рис. 87 приведена обоснованная в [Хорошевский, 2012] диаграмма количества публикаций, соответствующая тренду инновационной деятельности.

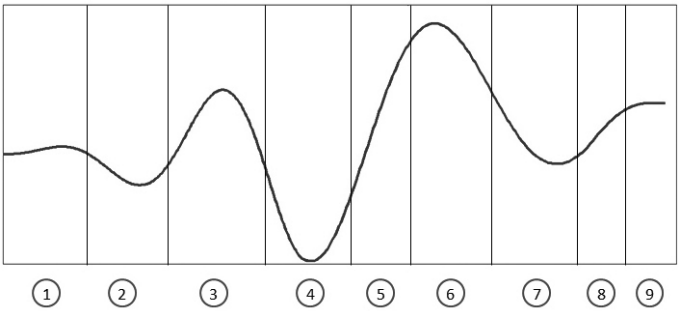


Рис. 87 – Обобщенная диаграмма, соответствующая всем этапам жизненного цикла информационных операций: 1 – фон; 2 – затишье; 3 – «артподготовка»; 4 – затишье; 5 – атака/триггер роста; 6 – пик завышенных ожиданий; 7 – утрата иллюзий; 8 – общественное осознание; 9 – продуктивность/фон

Объединяя графики, соответствующие началу информационной операции (рис. 52) и тренду инновационной деятельности (рис. 53), можно получить полный график, соответствующий отображению информационных операций в информационном пространстве (рис. 54).

Предложенные модели полностью соответствуют реальным данным, которые экстрагируются системами контент-мониторинга [Додонов, 2009], [Ландэ, 2007]. Поэтому приведенные зависимости могут быть использованы в качестве шаблонов для выявления информационных операций – как путем анализа ретроспективного фонда сетевых публикаций, так и путем оперативного мониторинга появления некоторых их признаков в реальном времени. Как известно, для выявления информационных операций следует внимательно следить за динамикой публикаций по целевой теме и, если есть возможность пользоваться доступными аналитическими средствами, средствами цифровой обработки данных и распознавания образов, например, вейвлет-анализом или полиномами Кунченко [Чертов, 2009].

В качестве примера, на рис. 88 показана динамика публикаций в RUNet тематических информационных потоков по запросам «Банки, Кипр», «Офшор», «Вирджинские острова» за март-апрель 2013 года, в период известных кризисных событий, полученная с помощью системы InfoStream. Как видно из рис. 55, пик публикаций, связанных с банковским кризисом на Кипре приходится на 17-18 марта 2013 года, в то время, как большинство публикаций по Вирджинским островам пришелось на 4–5 апреля, когда там, со значительно меньшими масштабами, стали проявляться события, подобные кипрским. При этом следует отметить слабую коррелированность динамики информационных потоков, связанных с Кипром и Вирджинскими островами. В этом случае коэффициент взаимной корреляции соответствующих числовых рядов составил всего 0,3. При этом отмечается высокий уровень взаимной корреляции рядов соответствующих тематикам «Офшор» и «Банки Кипра» (0,73), а также «Офшор» и «Вирджинские острова» (0,77).

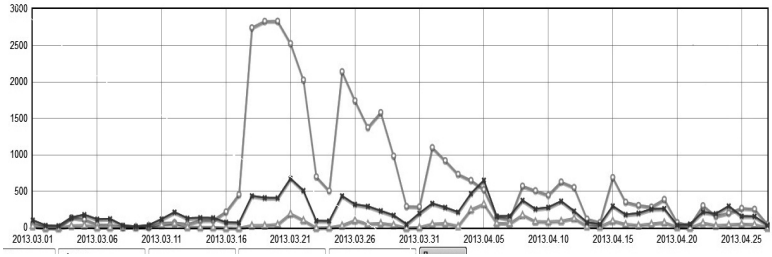
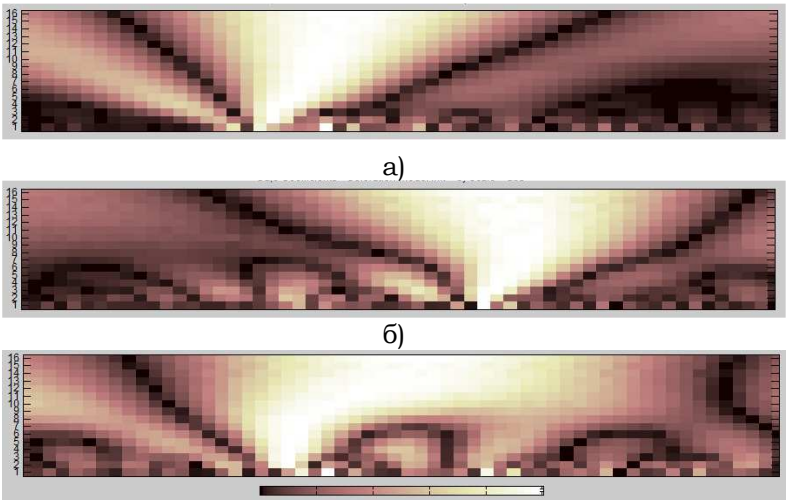


Рис. 88 – Диаграмма динамики тематических информационных потоков по запросам: o – «Банки Кипра»; Δ – «Вирджинские острова»; x – «Офшор»

По-видимому, проявления информационных операций в области оффшорных банков в данном случае лучше всего увидеть при анализе более общей тематики – «Офшоры». На графике соответствующего числового ряда четко видны две области локальных экстремумов, соответствующих кризисным ситуациям на Кипре и на Вирджинских островах, а также фазы, соответствующие «затишьям» и «артподготовкам».

Можно высказать предположение, что если динамика частного информационного потока в какой-то момент начинает существенно отличаться от динамики потока, соответствующего более общей тематике (как в рассматриваемом случае, «Банки Кипра» и «Офшор»), то возможно проявление признаков начала информационной операции, относящейся к узкой тематике.

При проведении вейвлет-анализа [Астафьева, 1996], [Buckheit, 1995] (рис. 89) было принято решение использования вейвлета «Мексиканская шляпа», как близкого по форме к диаграмме, приведенной на рис. 54. Рассматриваемые процессы четко просматриваются как на вейвлет-спектрограммах, так и на соответствующих им скелетонах (графиках линий экстремумов).



а)

б)

в)

Рис. 89 – Вейвлет-спектрограммы, соответствующие динамике тематических информационных потоков по запросам: а – «Банки Кипра»; б – «Вирджинские острова»; в – «Офшор»

Приведенные модели и методы пригодны для описания общих тенденций динамики информационных процессов, однако, проблема прогнозирования остается открытой. По-видимому, более реалистичные модели могут быть получены с учетом дополнительного набора факторов, большинство которых не воспроизводятся во времени. Вместе с тем, структура правил, лежащих в основе функционирования большинства из доступных моделей, позволяет вносить соответствующие коррективы, например, искусственно моделировать случайные отклонения.

Отметим, что воспроизведение результатов во времени является серьезной проблемой при моделировании информационных процессов и составляет основу научной методологии. В настоящее время только ретроспективный анализ уже реализованных информационных операций остается относительно надежным способом их верификации.

Естественно, на практике ориентация лишь на единственный тип источников может привести к дефициту информации, необходимой для принятия решений, неточностям, а порой – к дезинформированности. Лишь применение комплексных систем, базирующихся на использовании многочисленных источников и баз данных, наряду с приведенными выше возможностями системы контент-мониторинга, может гарантировать эффективную информационную поддержку при противодействии информационным операциям.

Выделенные образцы поведения рядов интенсивностей тематических публикаций могут рассматриваться как шаблоны (образцы) функциональной зависимости. Эти шаблоны можно взять в качестве единого базисного элемента некоторого линейного пространства, т.е. в качестве порождающего элемента *e* для моделирования с помощью полиномов Кунченко [Чертов, 2009].

Тогда можно построить полином P_n приближения n-го порядка к части выходного сигнала $f_s(e)$ как линейную комбинацию линейно-независимых преобразований $f_1(e), f_2(e),..., f_n(e)$:

$$P_n = \sum_{\substack{k=0,\\ k\neq s}}^{n} c_k f_k(e),$$

где коэффициенты c_k определяются из условия обеспечения минимума расстояния между строящимся полиномом и сигналом. Элемент c_0 определяется выражением:

$$c_0 = \frac{\langle f_s(e), f_0(e)\rangle - \sum\limits_{k=1,\ k\neq s}^{n} c_k \langle f_k(e), f_0(e)\rangle}{\langle f_0(e), f_0(e)\rangle},$$

а другие коэффициенты c_k – как решение системы линейных уравнений:

$$\sum_{k=1,\ k\neq s}^{n} c_k F_{i,k} = F_{i,s},\ i = 1,...,n,\ i \neq s,$$

где центрированные корреляты $F_{i,k}$ также рассчитываются с помощью соответствующих преобразований:

$$F_{i,k} = \langle f_i(e), f_k(e)\rangle - \frac{\langle f_i(e), f_0(e)\rangle \cdot \langle f_k(e), f_0(e)\rangle}{\langle f_0(e), f_0(e)\rangle}.$$

Числовой характеристикой, которую можно использовать в критериях качества сопоставления сигнала с выделенным шаблоном, т. е. как меру приближения полинома Кунченко P_n к сигналу $f_s(e)$, можно считать коэффициент эффективности d_n :

$$d_n = \frac{\sum_{k=1,\ k\neq s}^{n} c_k \left\langle f_k(e), f_s(e) \right\rangle}{\left\langle f_s(e), f_s(e) \right\rangle}.$$

Рассмотренный метод распознавания определенных образцов с помощью построения пространства с порождающим элементом и поиска коэффициентов соответствующего полинома Кунченко может быть использован в любой проблемной области, в которой можно априори во временном ряду выделить определенные характерные шаблоны.

Таким образом, построив типовые модели поведения рядов интенсивности тематических публикаций во время проведения информационных операций и сопоставив шаблоны, полученные на их основе, можно использовать метод на основе полиномов Кунченко для определения (и предупреждения) возможной информационной атаки.

Динамика тематических информационных потоков (ТИП) определяется комплексом как внутренних, так и внешних нелинейных механизмов, которые должны быть отражены при моделировании (возможно, в неявном виде). Зачастую удовлетворительным оказывается упрощенное понимание ТИП как некоторой зависимой от времени величины, поведение которой описывается в аналитическом виде нелинейными уравнениями. Сегодня при моделировании информационных потоков используются преимущественно аналитические нелинейные модели, применяются методы нелинейной динамики, теории клеточных автоматов, перколяции, самоорганизованной критичности [Ландэ, 2009], [Додонов, 2011].

Для анализа динамики реальных ТИП, и, соответственно, оценки их моделей необходимо каким-то образом получить соответствующую статистику, представленную в виде временных рядов. Динамику реальных ТИП, например, отображает мультиагентная

модель, в рамках которой отдельные документы ассоциируются с агентами, жизненный цикл которых – с жизненным циклом документов в информационном пространстве. Соответственно, все пространство мультиагентной модели ассоциируется с тематическим информационным потоком.

Предполагается, что в течение дискретных моментов времени происходит эволюция популяции агентов. При этом отдельные агенты могут:

1) «самозарождаться» (рождаться по причинам, возникающим вне рассматриваемого мультиагентного пространства);

2) «порождать» новых агентов;

3) «умирать» – исчезать из пространства агентов (соответст-вует утере актуальности документов);

4) получать ссылки от других агентов.

Каждый агент обладает «потенциалом», зависящим от его возраста (времени жизни на текущий момент – t), от авторитетности (ссылок, проставленных на него – ns) и плодовитости (количества порожденных непосредственно им агентов – k). Потенциал агента Pot определяется формулой:

$$Pot = \frac{1 + ns + k}{t}.$$

На рис. 90 приведен пример возможной динамики мультиагентой системы: процессы рождения новых агентов от существующих обозначены сплошными стрелками, процессы проставления ссылок на агентов представлены пунктирными стрелками, живые агенты – черными кругами, «мертвые» агенты к моменту t = 5 – незаполненными окружностями.

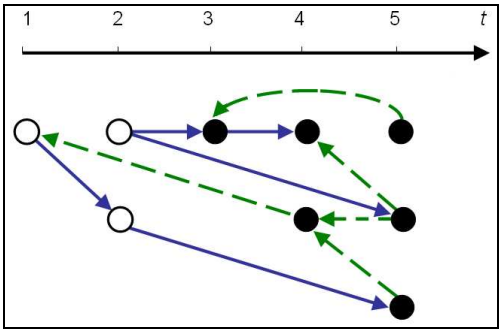


Рис. 90 – Фрагмент мультиагентного пространства

Итак, управляющие параметры модели следующие:

- вероятность «самозарождения» P_1;
- вероятность «рождения» от существующего: $P_2 \cdot Pot$;
- вероятность «смерти» агента: P_3 / Pot;
- вероятность ссылки на агента: $P_4 \cdot Pot$.

Варьирование этими четырьмя параметрами P_1, P_2, P_3 и P_4 позволили смоделировать типовые профили поведения ТИП.

На рис. 91 представлены результаты численного моделирования количества агентов (ось ординат на графике) в рассматриваемой мультиагентной системе в зависимости от количества тактов модели (ось абсцисс).

Рассматриваемая модель эволюции пространства агентов при различных значениях управляющих параметров согласуется с динамикой реальных тематических информационных потоков, определенных с помощью системы InfoStream.

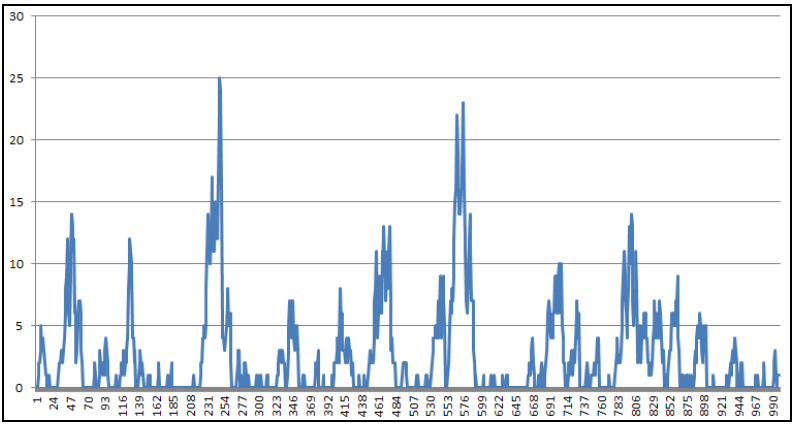


Рис. 91 – Динамика изменения количества агентов в модели

Наряду с исследованием огибающих динамики ТИП большой практический интерес представляет неравномерность, изрезанность соответствующих графиков, которая может свидетельствовать об отклонениях от естественной природы, информационных операциях, манипулировании [Горбулін, 2009]. В частности, для отображения неравномерностей во временном ряду использовался метод SCA (Smoothing, Cellular Automata) [Ландэ, 2013], основанный на учете аномальных значений и концепции одномерных клеточных автоматов. С помощью этого метода не детектируются абсолютные амплитудные всплески, однако он хорошо показал себя на «изрезанных» структурах данных, близких к фрактальным.

К таким данным относятся, в частности, временные ряды, связанные с объемами публикаций в веб-пространстве по определенным тематикам, которые рассматриваются ниже как иллюстрация метода.

В предлагаемой модели каждому значению исходного ряда измерений $x_0(t)$ (обозначим исходный

ряд, как $X_0 = \{x_0(t)\}$) соответствует одна клетка клеточного автомата. По ряду измерений строится сглаженный по приведенному ниже правилу ряд $X_1 = \{x_1(t)\}$. Затем ряду X_1 ставится в соответствие ряд X_2 (получаемый из X_1 по тому же алгоритму сглаживания) и т.д. Правило сглаживания пиков заключается в том, что значения, которые принимают элементы рядов измерений $x_k(t) \in X_k$ (k – шаг сглаживания, t – номер элемента ряда измерений) составляют:

$$x_k(t) = \begin{cases} x_{k-1}(t), \ \ if \ \ x_{k-1}(t) \le \dfrac{x_{k-1}(t-1) + x_{k-1}(t+1)}{2}, \\ \dfrac{x_{k-1}(t-1) + x_{k-1}(t+1)}{2}, \ \ if \ \ x_{k-1}(t) > \dfrac{x_{k-1}(t-1) + x_{k-1}(t+1)}{2}. \end{cases}$$

Цвет клетки с номером t одномерной клеточной структуры, соответствующей X_k, белый, если $x_k(t)$ совпадает с $x_{k-1}(t)$, в противном случае – черный. Таким образом, каждой клетке соответствует значение $x_k(t)$ и значение ее цвета. (Необходимо отметить, что такую систему нельзя считать каноническим клеточным автоматом, так как в общем случае клеткам может соответствовать бесконечное множество значений $x_k(t)$ и два значения цвета).

Рассмотрим результаты выполнения данного алгоритма для простейших структур, которые, как показывает практика, охватывают все возможные варианты визуализации.

Очевидно, если значения ряда измерений в рассматриваемой зоне представляют собой вогнутое (выпуклое вниз множество), то сразу же на первой итерации получается $\forall t: \ x_1(t) = x_0(t)$ и выполнение алгоритма прерывается.

Если область значений представляют собой

выпуклое вверх множество, то визуальное представление клеточных автоматов принимает вид сплошной черной полосы (рис. 92: вертикальная ось – номер шага итерации, а горизонтальная ось – номер элемента ряда измерений). Единичные всплески значений в исходном ряде измерений (рис. 93 *а*) и области изрезанности (рис. 93 *б*) могут вызывать появление структур типа «шахматной доски».

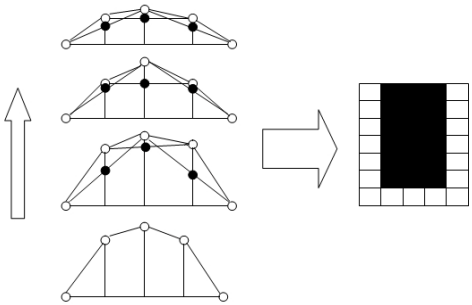


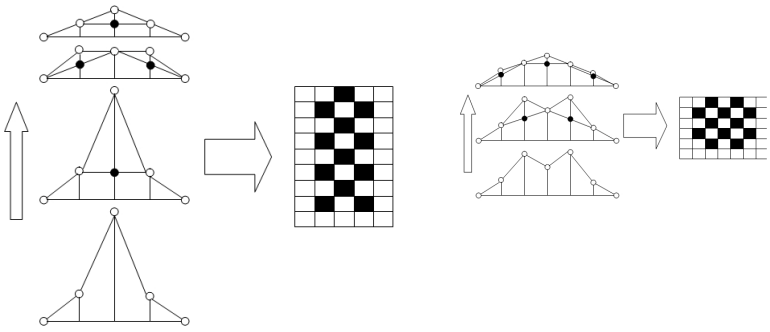
Рис. 92 – Выпуклое вверх множество точек

Кроме того, диаграммы, формируемые в результате визуализации в соответствии с предложенным алгоритмом, позволяют выявлять периодические составляющие.

Предложенный метод SCA является относительно простым в программной реализации и линейным по сложности, так как базируется на алгоритме сглаживания пиков и концепции клеточных автоматов. Он позволяет визуально выявлять единичные и нерегулярные «всплески», резкие колебания, скачки значений, зоны нестабильности количественных показателей в разные периоды времени. Метод SCA испытывался при анализе временных рядов, связанных с объемами публикаций в веб-пространстве по определенным темам.

На диаграммах, формируемых в соответствии с методом SCA, выпуклое вверх множество принимает вид сплошной черной полосы, выпуклое вниз множество – белой полосы, а области изрезанности,

нестабильности могут вызывать появление «клетчатых» структур.



а) б)

Рис. 93 – Появление структур типа «шахматной доски»

Отображение реального временного ряда измерений, соответствующего посуточным объемам публикаций в веб-пространстве по некоторой заданной теме (точки ряда – объемы публикаций за сутки) с помощью метода SCA представлено на рис. 94.

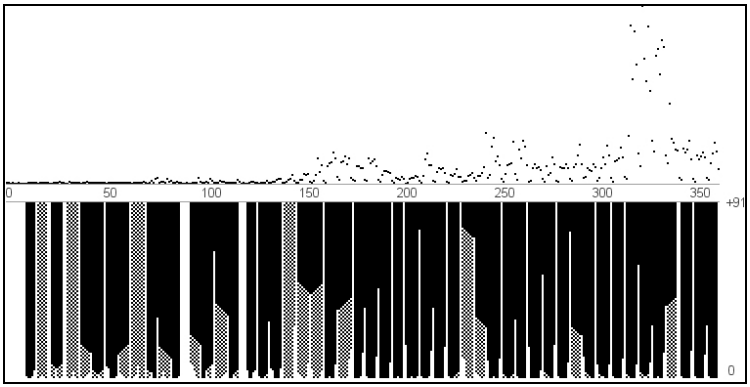


Рис. 94 – Отображение реальной динамики публикаций в течение года с помощью метода SCA

Здесь четко отслеживаются недельные периодичности ТИП (минимумы – праздники, субботы и

воскресенья), а также области неравномерности, резких колебаний объемов публикаций, свойственных ТИП в пред- и посткризисные периоды.

На рис. 95 представлена SCA-визуализация динамики количества агентов (документов) – результатов мультиагентного моделирования.

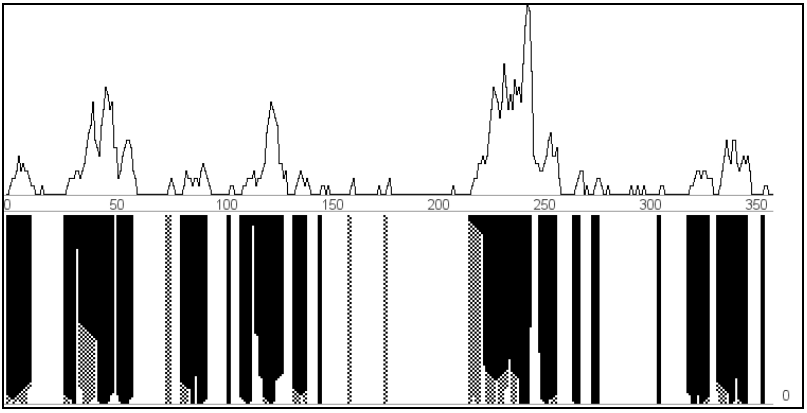


Рис. 95 – Отображение результатов мультиагентного моделирования с помощью метода SCA

Следует отметить, что предлагаемая модель:

1) не учитывает конкуренции агентов внутри пространства агентов (предполагается только сотрудничество путем проставления ссылок и порождения новых агентов);

2) конкуренция разных тематических информационных потоков учитывается лишь неявно, как причина, обуславливающая параметры функционирования рассматриваемой мультиагентной системы.

В предложенной модели учитывается общеизвестная практика проведения информационных кампаний в социальных сетях, заключающаяся в регистрации большого числа аккаунтов-роботов (роя), от имени которых проставляются ссылки (лайки) на

материалы, публикуемые от имени аккаунтов из того же роя и на целевые документы.

Естественно, на практике ориентация лишь на единственный тип источников и математических моделей может привести к дефициту информации, необходимой для принятия решений, неточностям, а порой – к дезинформированности. Лишь применение комплексных систем, базирующихся на использовании многочисленных источников, баз данных, математических моделей, наряду с приведенными выше возможностями систем контент-мониторинга может гарантировать эффективную информационную поддержку при противодействии информационным операциям.

5.6. Анализ динамики событий

Для эффективного проведения информационно-аналитических исследований на основе анализа контента сети Интернет (а точнее ее веб-ресурсов) авторами предлагается последовательность шагов, этапов обработки информации, каждый из которых сам по себе обеспечивает получение продукта. Совокупность таких этапов, базирующихся на использовании необходимых и доступных инструментальных средств, специальных приемов, можно рассматривать как методику, процедуру проведения действий, нацеленных на получение аналитических материалов, которые могут использоватьсяя для поддержки принятия решений.

Любая методика рассчитана на решение конкретных задач. При проведении информационно-аналитических исследований на базе интернет-контента к таким задачам можно отнести:

- Нахождение релевантных публикаций по тематике.
- Определение динамики тематических публикаций.

- Определение критических точек в динамике тематических публикаций.
- Выявление объектов мониторинга
- Выявление и визуализация взаимосвязей событий и объектов мониторинга, а также объектов мониторинга между собой.
- Прогнозирование развития событий

Этапы информационно-аналитического исследования

В соответствии с приведенными выше задачами предлагается выделить следующие этапы информационно-аналитического исследования:

- Выбор системы интеграции интернет-документов.
- Формирование запроса в среде выбранной системы. Нахождение тематических публикаций по запросу с помощью систем контент-мониторинга.
- Определение динамики тематических публикаций по запросу.
- Определение критических точек в динамике тематических публикаций.
- Определение основных событий в критических точках.
- Выявление объектов мониторинга.
- Выявление и визуализация взаимосвязей.
- Прогноз развития событий

Рассмотрим эти этапы подробно.

Выбор системы интеграции интернет-документов

Для получения репрезентативной информации об объекте исследования необходимо воспользоваться поисковой системой с аналитическими функциями, охватывающей достаточный объем информации, относящейся к исследуемому объекту/событию. Для

анализа динамики информационных потоков необходимо каким-то образом получить соответствующую статистику, представленную в виде временных рядов. Многие современные информационно-аналитические системы содержат в своем составе средства отображения статистики вхождения в базы данных понятий, соответствующих пользовательским запросам. В настоящее время существует несколько открытых информационных сервисов, в рамках которых можно наблюдать временную динамику объемов публикаций по тематикам, определяемым запросами. Так Google books Ngram Viewer (http://ngrams.googlelabs.com/), предоставляет визуализацию динамики количества книг, в которых упоминаются слова. Сервис «Яндекс пульс блогосферы» (http://blogs.yandex.ru/pulse/) также позволял отображать динамику публикаций в блогах, содержащих заданные пользователем ключевые слова, однако был закрыт ввиду малой посещаемости. Сегодня этот сервис доступен лишь по специальному разрешению компании «Яндекс». На сайте Национального корпуса русского языка (НКРЯ) в бета-режиме запущен сервис *N*-грамм (http://www.ruscorpora.ru/ngram.html), близкий по функциональности сервису Google books Ngram Viewer.

Безусловно, самыми эффективными для решения задач анализа динамического контента являются специализированные системы интеграции сетевого контента. В частности, в рамках описываемых исследований авторами использовалась система контент-мониторинга веб-ресурсов InfoStream (www.infostream.ua), реализующая необходимую функциональность и охватывающая около 100 тыс. документов в сутки с 7000 веб-сайтов.

Формирование запроса в среде выбранной системы

Массив тематических документов (тема – события, связанные с Евромайданом в Киеве 2013-2014 гг.) выбирается с помощью системы InfoStream путем ввода

запроса на языке данной системы:

(майдан|евромайдан)&(избиен|разгон|штурм|беркут|молотов|титушк|погиб)&lang.RUS,

по которому в период с ноября 2013 года по март 2014 года было опубликовано свыше 200 тысяч *тематических публикаций* на веб-сайтах (рис. 96).

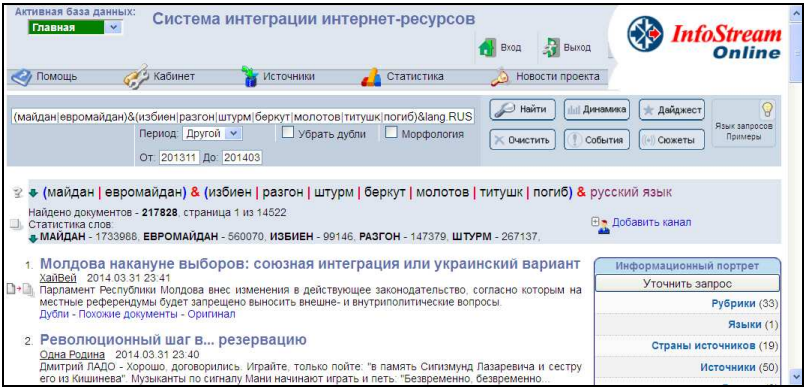


Рис. 96 – Поисковый интерфейс системы InfoStream

Система InfoStream обеспечивает поиск, а также просмотр списка и полных текстов релевантных документов.

Определение динамики тематических публикаций по запросу

Режим динамики событий системы интеграции интернет-ресурсов позволяет получить данные о количестве публикаций по заданному запросу за указанный промежуток времени. Эти данные могут быть загружены в настольную систему обработки данных и отображаются в виде графика (рис. 97). В интерфейсе пользователя обеспечивается переход к просмотру релевантных документов по выбранной дате.

После этого данные временной динамики за

каждые сутки нормируются, т.е. в среде системы Excel формируется временной ряд, содержащий относительные значения, равные отношению количества тематических сообщений к общему потоку сообщений за сутки.

Рис. 97 – Режим «Динамика событий» системы интеграции

Это, в частности, позволяет избавиться от недельной периодичности в количестве тематических публикаций. Затем происходит переход к процедуре определения критических точек в данном временном ряде.

Критические точки как локальные максимумы временного ряда динамики публикаций можно определить, например, визуально по графику, представленному на рис. 97. Вместе с тем, существуют несколько научно-обоснованных методик, базирующихся на методах цифровой обработки сигналов.

Цикл информационных операций

В результате анализа многочисленных диаграмм поведения ТИП были выявлены наиболее типичные, базовые профили их поведения [Додонов, 2013]. Предложенные модели полностью соответствуют реальным данным, которые экстрагируются системами

контент-мониторинга. Поэтому приведенные зависимости могут быть использованы как шаблоны, например, для выявления информационных операций – как путем анализа ретроспективного фонда сетевых публикаций, так и оперативного мониторинга появления некоторых их признаков в реальном времени.

В частности, для выявления информационных операций [Горбулін, 2009] следует внимательно следить за динамикой публикаций по целевой теме и, если есть возможность, пользоваться доступными аналитическими средствами, средствами цифровой обработки данных и распознавания образов, например, вейвлет-анализом.

Выше, на рис. 87 была приведена обобщенная диаграмма, соответствующая всем этапам жизненного цикла информационных операций, обоснованная в [Ланде, 2008].

Для выявления степени «близости» фрагментов исследуемого временного ряда приведенной диаграммы в различных масштабах предлагается использовать так называемый «вейвлет-анализ» [Астафьева, 1996], который в настоящее время нашел широке применение как в естественных науках, так и в социологии [Давыдов, 2008].

Главная идея вейвлет-преобразования заключается в том, что нестационарный временной ряд разделяется на отдельные промежутки (так называемые «окна наблюдения») и на каждом из них вычисляется величина, показывающая степень близости закономерностей исследуемых данных с разными сдвигами некоторого вейвлета (специальной функции) в разных масштабах. Вейвлет-преобразование генерирует набор коэффициентов, которые являются функциями двух переменных: времени и частоты, и потому образовывают поверхность в трехмерном пространстве.

Непрерывное вейвлет-преобразование для функции $f(t)$ строится с помощью непрерывных

масштабных преобразований и переносов выбранного вейвлета $\psi(t)$ с произвольными значениями масштабного коэффициента a и параметра сдвига b:

$$W(a,b) = (f(t),\psi(t)) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} f(t)\psi^*\left(\frac{t-b}{a}\right)dt .$$

Полученные вейвлет-коэффициенты можно представить в графическом виде, если по одной оси отложить сдвиг вейвлета (ось времени), а по другой – масштабы (ось масштабов), и раскрасить точки полученной схемы в зависимости от величины соответствующих коэффициентов (чем больше коэффициент, тем ярче цвета).

Эти коэффициенты показывают, насколько поведение процесса в данной точке аналогично вейвлету в данном масштабе. Чем ближе от анализируемой зависимости в окрестности данной точки к виду вейвлета, тем большую абсолютную величину имеет соответствующий коэффициент. Применение этих операций, с учетом свойства локальности вейвлета в частотно-временной области, позволяет анализировать данные на разных масштабах и точно определять положение их характерных особенностей во времени.

На скейлограмме видны все характерные особенности исходного ряда: масштаб и интенсивность периодических изменений, направление и значение трендов, наличие, расположение и продолжительность локальных особенностей.

В работе [Додонов, 2013] показано, что вейвлеты «мексиканская шляпа» и Морле (рис. 98) наиболее точно отражает динамику информационных операций, результаты применения этого вейвлета приведены на рис. 99, благодаря чему выбраны три даты (2013.11.30, 2014.01.22, 2014.02.19), соответствующие критическим точкам исследуемого процесса.

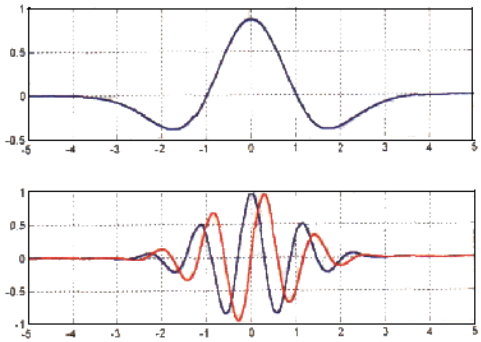


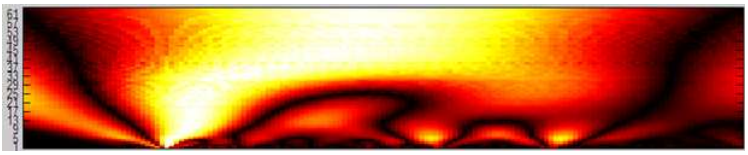
Рис. 98. Вейвлеты mexh, Морле

Следует отметить, что инструменты построения вейвлет-спектограмм доступны как в ряде пакетов математических программ, например, в Matlab, так и через Интернет в режиме онлайн (http://ion.researchsystems.com/cgi-bin/ion-p?page=wavelet.ion).

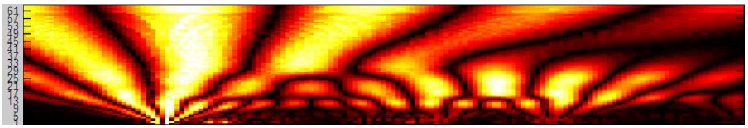
Определение основных событий в критических точках

После определения критических точек с помощью системы контент-мониторинга выполняется построение основных сюжетных цепочек из сообщений, соответствующих запросу за выбранные даты. Таким образом определяются основные события по указанным датам (рис. 100).

Для последующего анализа отбираются три массива сообщений, соответствующие трем выбранным датам, объекты из которых (в простейшем случае – персоны и веб-источники) могут рассматриваться как объекты мониторинга.



а



б

Рис. 99 – Вейвлет-спектограммы исследуемого временного ряда (а – «мексиканская шляпа», б – вейвлет Морле)

Рис. 100 – Основные сюжетные цепочки по выбранным датам

Выявление объектов мониторинга

С помощью методов экстрагирования фактических данных, применяемых в системах интеграции интернет-ресурсов, в интерфейсе пользователя формируются так называемые «информационные портреты», охватывающие списки персон, топонимов, языков, компаний и т.п., содержащиеся в документах, релевантных некоторому заданному запросу.

В рассматриваемом случае из «информационного портрета», соответствующего тематическому запросу, выбираются наиболее упоминаемые персоны и/или веб-ресурсы за выбранные даты (рис. 101 и 102). Эти списки могут агрегироваться, в результате чего возможно определение взаимосвязей событий и объектов (рис. 103).

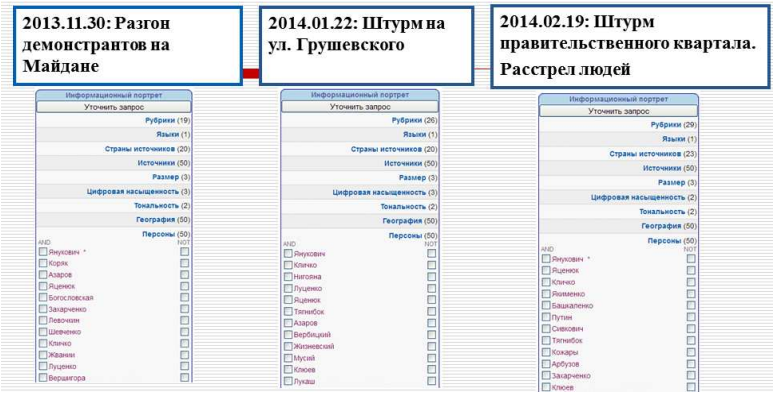


Рис. 101 – Списки «доминирующих» персон

В качестве системы визуализации авторами выбрана система анализа и отображения сетей Gephi (www.gephi.org).

Эти же данные позволяют выявлять взаимосвязи между объектами, например, между указанными аналитиками веб-ресурсами и персонами (рис. 104).

На рис. 103 можно видеть, что каждому массиву (узлы, идентифициорованные датами) соответствуют объекты. При этом в центральной части сети располагаются объекты, общие для нескольких событий (О-зона), а «гребешки» на периферии соответствуют специальным объектам, отражающим специфику конкретных событий (С-зоны).

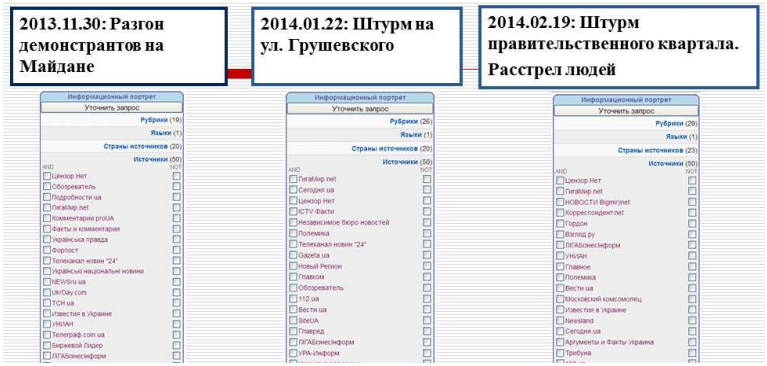


Рис. 102 – Списки «доминирующих» веб-ресурсов

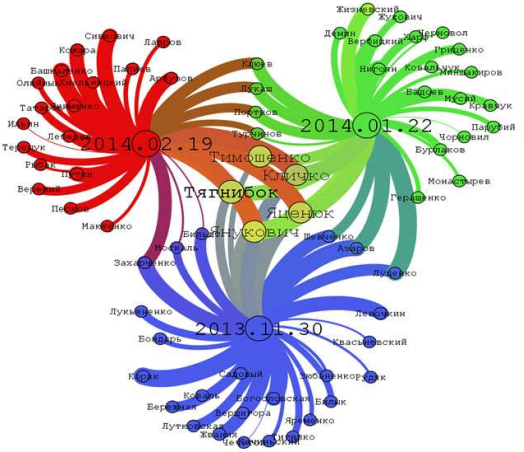


Рис. 103 – Пример визуализации связей событий и объектов

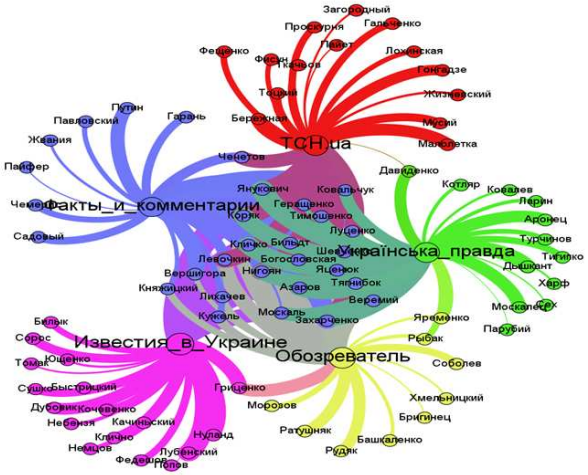


Рис. 104 – Визуализация взаимосвязей веб-ресурсов и персон

Также можно предложить критерий релевантности события, связанного с конкретной датой, общей тематике: чем большая часть объектов из него попадает в О-зону, тем он более релевантен тематике. Формально значение этого критерия $k_{i,N}$ для сюжета i тематики s может быть записано следующим образом:

$$k_{i,N} = \frac{\left| T_{i,N} \bigcap T_{s,N} \right|}{N},$$

где N – количество объектов, $T_{i,N}$ – множество значимых объектов события i, – $T_{s,N}$ – множество значимых объектов для всей тематики.

Подход к прогнозу: R/S-анализ

Для решения задач прогнозирования перспективным представляется применение теории фракталов при анализе информационного пространства. Фрактальный анализ самоподобия информационных массивов может рассматриваться как

технология, предназначенная для осуществления аналитических исследований с элементами прогнозирования, пригодная к экстраполяции полученных зависимостей.

Важнейшей характеристикой рядов, которые имеют хаотичное поведение, является, как известно, показатель Херста (H), определяемый в результате так называемого R/S-анализа [Федер, 1991]. Этот показатель базируется на анализе нормированного разброса – отношения разброса значений исследуемого ряда R к стандартному отклонению S.

Достаточно часто, когда соотношение R/S имеет постоянный тренд, можно говорить о соотношении:

$$R/S = (N/2)^H,$$

где H – показатель Херста, который для достаточно широкого класса рядов связан с хаусдорфовой (фрактальной) размерностью D простой формулой: $D+H=2$.

На рис. 105 представлено соотношения R/S для рассматриваемого в этой работе временного ряда. Как можно видеть, кривая нормированного размаха удовлетворительно аппроксимируется прямой в двойном логарифмическом масштабе. Численные значения H характеризуют разные типы коррелированной динамики (персистентности). При $H=0,5$ наблюдается некоррелированное поведение значений ряда, а значение 0,5 < H < 1 соответствует степени автокорреляции ряда.

Как можно видеть, значение показателя Херста для исследуемого информационного потока соответствует величине 0,81, что подтверждает предположение о самоподобии и итеративности рассматриваемых процессов в информационном пространстве. Это означает, что общая информационная напряженность остается на большом уровне, как только исчезает «шлейф» одного сюжета по выбранной тематике, ему на

смену возникает новый сюжет, т. е. его поведение в дальнейшем будет близко к предшествующему поведению.

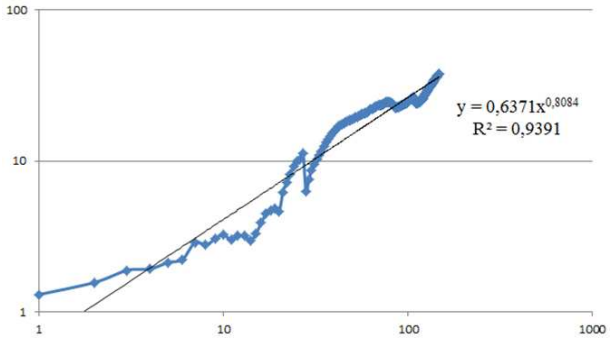


Рис. 105 – Кривая *R/S* в двойной логарифмической шкале

Таким образом, представлена методика аналитического исследования, которая базируется на использовании современных инструментальных средствах анализа и визуализации информационных потоков и временных рядов. Предложенную методику можно использовать в качестве основы для проведения аналитической и прогнозной деятельности на основе исследования контента современных компьютерных сетей.

5.7. Противодействие информационным операциям

Рассмотренные практические примеры позволили выработать некоторую общую методику проведения оборонительной информационной операции с использованием системы контент-мониторинга веб-ресурсов. Допустим, объектом агрессивной информационной операции является компания «АБВ». Предлагается такие 12 шагов противодействия:

1) сбор информации с публикациями в «чужих» (не имеющих отношения к «АБВ», неаффилированных) СМИ о компании;

2) построение графика – динамики появления сообщений о компании «АБВ» в сетевых СМИ;

3) анализ динамики с ретроспективой в 6–12 месяцев с помощью методов анализа временных рядов. После этого анализируется контент публикаций в пороговых точках, определяются моменты, длительность, периодичность воздействия, привязка моментов воздействия к другим событиям из области интереса объекта;

4) определение источников, публикующих наибольшее количество негатива (публикаций с отрицательной тональностью) о компании «АБВ»;

5) определение «первоисточников» публикаций в СМИ – тех источников, которые первыми опубликовали негативную информацию;

6) определение вероятных «заказчиков» – владельцев или лиц, влияющих на издательскую политику отдельных СМИ;

7) определение сфер общих интересов компании «АБВ» и потенциальных «заказчиков» (путем выявления общих информационных характеристик – пересечений «информационных портретов» системы InfoStream, строящихся для объекта и «заказчика»), ранжирование потенциальных «заказчиков» по их интересам;

8) определение критериев информационных воздействий на основе самых рейтинговых интересов;

9) моделирование информационных воздействий, для чего находятся связи «заказчика» – наиболее связанные с ним персоны и организации, анализируется динамика воздействия со стороны заказчика и строится прогноз этой динамики, анализируется контент публикаций в пороговых точках кривой динамики – определяются критичные точки воздействия;

10) прогнозируются дальнейшие шаги воздействия путем анализа аналогичной динамики публикаций для

других компаний в ретроспективной базе данных системы InfoStream;

11) с учетом реалий и публикаций из ретроспективной базы данных оцениваются вероятные последствия;

12) организуется информационное (и не только) противодействие. Примеры публикаций в контексте противодействия находятся в ретроспективной базе данных.

ЗАКЛЮЧЕНИЕ

Аналитическая деятельность, моделирование – необходимые компоненты как изучения, оценки, планирования и прогнозирования процессов, процедур в любых областях, так и для изучения их последствий.

Информационные технологии, необходимые для проведения сетевой аналитики, охватывают средства анализа информационного наполнения современных компьютерных сетей, динамических информациооных массивов сверхбольшого объема, потоков информации, непрерывно появляющейся в глобальной сетевой среде.

Одной из базовых компонент технологий поддержки аналитических исследований в сетевой среде является моделирование, выступающее, кроме того, фундаментом таких направлений как проектирование и прогнозирование. Рассмотреные в работе подходы позволяют строить абстрактные модели, которые в определенном приближении позволяют описывать информационные процессы, процессы информационного влияния. Подобные модели пригодны для описания общих тенденций в динамике информационных систем. Более реалистичные модели могут быть получены при учете большого количества факторов, большинство из которых не воссоздаются во времени. В тоже время повторяемость явлений, которые моделируются, является основой научной методологии. Поэтому воссоздание результатов во времени является самой серьезной проблемой при моделировании информационных процессов, в частности, информационных операций. В настоящее время лишь ретроспективный анализ уже реализованных процессов является надежным способом верификации результатов. Современные подходы позволяют применять для моделирования даже общественных и информационных систем методы, апробированные в первую очередь в естественных науках. Однако, следует отметить, что подходы, которые базируются на применении точных методов и математическом формализме, а также методов компьютерного

моделирования, в действительности, могут давать преимущественно качественные выводы, что обусловливается многопараметричностью реальных моделей. Вместе с тем, даже такие результаты могут объяснить реальность во многих случаях лучше, чем традиционные подходы.

Модели, алгоритмы, инструментарий поддержки аналитической деятельности, которые рассматриваются в данной монографии, сегодня выступают не только в качестве демонстрационной основы для объяснения реально происходящих событий и процессов, но и как необходимые компоненты при их планировании и прогнозировании. При этом продвижение в освоении современного информационного пространства невозможно без общих представлений о структуре и свойствах динамики сетевых информационных процессов, что в свою очередь требует выявления и учета их устойчивых закономерностей.

Эффективность информационно-аналитической работы в любой области зависит как от информационных технологий, уровня сетевой инфраструктуры, мощности соответствующих информационных потоков, программного инструментария, так и от стемени проработки теоретических основ аналитической деятельности.

Построение адекватных моделей сетевого информационного пространства, информационных потоков, принципов навигации и информационного поиска в этом пространстве, очевидно, должны базироваться на научных достижениях в областях искусственного интеллекта, социологии, концепции сложных сетей, теории принятия решений, компьютерной лингвистики.

ЛИТЕРАТУРА

[Albert, 1999] *Albert R., Jeong H., Barabasi A.-L.* Diameter of the world wide web // Nature (London), 1999. – 401. – P 130.

[Amsler, 1982] *Amsler R.A.* Computational lexicology: A research program. In American Federated Information Processing Societies Conference Proceedings. – National Computer Conference, 1982. – P 657-663.

[Blekanov, 2014] *Blekanov I.S., Sergeev S.L., Maksimov A.I.* Analysis of the topology of large Web segments using Broder's bow-tie model // Life Science Journal, 2014. – № 11. – (6s). – 258-261.

[Bhargava, 1993] *Bhargava S.C., Kumar A., Mukherjee A.* A stochastic cellular automata model of innovation diffusion // Technological forecasting and social change, 1993. – 44. – № 1. – P. 87-97.

[Bourdaillet, 2007] *Bourdaillet J.* Alignment of Noisy Unstructured Text Data // IJCAI-2007 Workshop on Analytics for Noisy Unstructured Text Data. Hyderabad, India. – January 8, 2007. P. 139-146.

[Bradford, 1934] *Bradford S.C.* Sources of Information on Specific Subjects // Engineering: An Illustrated Weekly Journal (London), 137, 1934 (26 January). – P. 85-86.

[Broder, 1997] *Broder A., Glassman S.C., Manasse M.S.* Syntactic Clustering of the Веб // WWW6, 1997.

[Broder, 2000] *Broder A., Kumar R, Maghoul F etc.* Graph structure in the Web // Proceedings of the 9th international World Wide Web conference on Computer networks: the international journal of computer and telecommunications networking. Amsterdam, The Netherlands, 2000. – P. 309-320.

[*Broder, 2000-1*] *Broder A.* Identifying and Filtering Near-Duplicate Documents, COM'00 // Proceedings of the 11th Annual Symposium on Combinatorial Pattern Matching. –

2000. – P. 1-10.

[Bruton, 1960] *Bruton R., Kebler R.* The half-life of some scientific and technical literature. // Am. document., 1960. – **11**. – № 1. – P. 18-22.

[Buckheit, 1995] *Buckheit J., Donoho D.* Wavelab and reproducible research // Stanford University Technical Report 474: Wavelets and Statistics Lecture Notes, 1995. – 27 p.

[Burke, 2001] *Burke M.M.* Knowledge Operations: above and beyond Information Operations // 6th International Command and Control Research and Technology, June 19 – 21, 2001. – 16 p.

[Caldeira, 2005] *Caldeira S. M. G., Petit Lobao T. C., Andrade R. F. S., Neme A., Miranda J. G. V.* The network of concepts in written texts // Arxiv preprint physics/0508066. –2005.

[Chowdhury, 2002] *Chowdhury A., Frieder O. etc.* Collection statistics for fast duplicate document detection // ACM Transactions on Information Systems (TOIS), April 2002. – 20, Issue 2. – P. 171-191.

[Document, 2008] Document management – Portable document format – Part 1: PDF 1.7 // Adobe Systems Inc. – 2008. – 756 p.

[Boyle, 2009] *Boyle A.* Net not as interconnected as you think. – Режим доступа: //www.news.zdnet.com/2100-9595_22-502388.html

[Clauset, 2008] 13. *Clauset A., Moore C., Newman M.E.J.* Hierarchical structure and the prediction of missing links in networks // Nature 453, 98-101 (1 May 2008).

[Devine, 2013] Devine J., Egger-Sider F. Going Beyond Google Again: Strategies for Using and Teaching the Invisible Web. – Facet Publishing, 2013. – 224 p.

[DoD, 2003] Information operations roadmap – DoD US, 30 october 2003. – 78 p.

[Dorogovtsev, 2001] *Dorogovtsev S.N., Mendes J. F. F.* Language as an evolving word web // Proc. R. Soc. Lond., 2001. – **B 268**, 2603.

[Dorogovtsev, 2003] *Dorogovtsev S.N., Mendes J.F.F.* Evolution of networks: from biological networks to the Internet and WWW. – Oxford University Press, 2003.

[Feigenbaum, 1978] *Feigenbaum M.J.* (1978) Quantitative universality for a class of nonlinear transformations. J. Stat. Phys., 1978. – **19**. 25-52.

[Fellbaum, 2005] Fellbaum C. WordNet: An Electronic Lexical Database. – MIT Press, 2005. – 425 p.

[Ferrer-i-Cancho, 2001] *Ferrer-i-Cancho R.*, *Sole R. V.* The small world of human language // Proc. R. Soc. Lond., 2001. – **B 268**. – P. 2261.

[Ferrer-i-Cancho, 2004] *Ferrer-i-Cancho R.*, *Sole R.V., Kohler R.* Patterns in syntactic dependency networks // hys. Rev., 2004. – **E 69**, 051915.

[Ferrer-i-Cancho, 2005] *Ferrer-i-Cancho R.* The variation of Zipf's law in human language. // Phys. Rev., 2005. – **E 70**, 056135.

[Frantz, 2005] *Frantz T., Carley K.M.* A formal characterization of cellular networks // Carnegie Mellon University School of Computer Science Institute for Software Research International, Tech. Rep. CMU-ISRI-05-109, 2005.

[Giora, 1983] *Giora R.* Segmentation and Segment Cohesion: On the Thematic Organization of the Text // Text. An Interdisciplinary Journal for the Study of Discourse Amsterdam. – **3**. – № 2. – P. 155-181 (1983).

[Gutin, 2011] Gutin G., Mansour T., Severini S. A characterization of horizontal visibility graphs and combinatoris on words // Physica A, 2011. – 390 – P. 2421-2428.

[Haken, 1964] *Haken H.* Statistische nichtlineare Theorie des Laserlichts. Z. Physik. B. 181, 1964. Z.96.

[Haken, 1977] *Haken H.* Synergetics — An Introduction; Noncquilibrium Phase Transitions and Self-Organization in Physics, Chemistry and Biology. Springer-Verlag, 1977. (Рус. пер.: Хакен Г. Синергетика. М., 1980.)

[Hawking, 1999] *Hawking D., Thistlewaite P.* Methods for Information Server Selection. ACM Transactions on Information Systems, 17(1), January 1999.

[He, 2007] He B., Patel M., Zhang Z., Chang K. C. C. Accessing the Deep Web: A Survey // Communications of the ACM (CACM), 50(5):94-101, 2007.

[Hill, 2000] *Hill J.M.D., Surdu J.R., Ragsdale D.J., Schafer J.H.* Anticipatory planning in information operations // Systems, Man, and Cybernetics, 2000 IEEE International Conference, 2000. – **4**. – P. 2350-2355.

[Hutchins, 2005] *Hutchins W.J.* Current commercial machine translation systems and computer-based translation tools: system types and their uses // International Journal of Translation, 2005. – **17**. – № 1-2. – P. 5-38.

[Hutchins, 2007] *Hutchins W.J.* Machine translation: a concise history // To be published in Computer aided translation: Theory and practice, ed. Chan Sin Wai. Chinese University of Hong Kong, 2007.

[Ilyinsky, 2002] *Ilyinsky S., Kuzmin M., Melkov A., Segalovich I.* An efficient method to detect duplicates of Веб documents with the use of inverted index // WWW-2002 – Eleventh Intern. World Wide Веб Conference. URL: http://www2002.org/CDROM/poster/187/.

[Kacperski, 2000] *Kacperski K., Holyst J.A. Physica A.* Phase transitions as a persistent feature of groups with leaders in models of opinion formation // Statistical Mechanics and its Applications, 2000. – 287, Issues 3-4. – P 631-643.

[*Karp,1987*] *Karp R.M., Rabin M.O.* Efficient randomized pattern-matching algorithms // IBM Journal of Research and Development. – 31 (2), March 1987. 249-260.

[Kleinberg, 1998] *Kleinberg J.* Authoritative sources in a hyperlinked environment // In Processing of ACM-SIAM Symposium on Discrete Algorithms, 1998, 46(5):604-632.

[Kowalczyk, 2014] Kowalczyk M., Buxmann P. Big Data and Information Processing in Organizational Decision Processes// Business & Information Systems Engineering, 2014. – 6. – № 5. – P. 267-278.

[Lande, 2007] *Lande D., Braichevski S., Busch D.* Informationsfluesse im Internet // IWP -Information Wissenschaft & Praxis, 2007. – **5**. – № 59 – P. 277-284.

[Lande, 2008] *Lande D.V., Zhygalo V.V.* About the creation of a parallel bilingual corpora of web-publications // ePreprint Arxiv (0807.0311).

[Lande, 2013] *Lande D.V., Snarskii A.A., Yagunova E.V., Pronoza E.V.* The Use of Horizontal Visibility Graphs to Identify the Words that Define the Informational Structure of a Text // 12th Mexican International Conference on Artificial Intelligence, 2013. – P. 209-215.

[Lande, 2014] *Ландэ Д.В., Снарский А.А.* Подход к созданию терминологических онтологий // Онтология проектирования, 2014. – N 2(12). – С. 83-91.

[Lasswell, 1948] *Lasswell H.D.* The structure and function of communication in society // The Communication of Ideas. / Ed.: L. Bryson. – New York: Harper and Brothers, 1948.

[Latane, 1981] *Latane B.* The psychology of social impact // American Psychologist, 1981. – **33**. – P. 343-356.

[Latane, 1997] *Latane B., Nowak A.* Causes of polarization and clustering in social groups // Progress in communication sciences, 1997. – **13**. – P. 43-75.

[Lawrence, 1998] *Lawrence S., Lee G.C.* Searching the World Wide Web // Science, 280, April 1998.

[Lawrence, 1998] *Lawrence S., Lee G.C.* Inquirus, the NECi Meta Search Engine // Seventh International World Wide Web Conference, 1998.

[Lempel, 2000] *Lempel R., Moran S.* The stochastic approach for link-structure analysis (SALSA) and the TKC effect // In Proceedings of the 9th International World Wide Web Conference, Amsterdam, The Netherlands, 2000. – P. 387-401.

[Lewenstein, 1993] *Lewenstein M., Nowak A., Latane B.* Statistical mechanics of social impact // Physical Review, 1993. – A, 45. – P. 763-776.

[Liu, 2000] *Liu K., Yu C., Meng W.* Discovering the Representative of a Search Engine // Technical Report, DePaul University, 2000.

[Liu, 2001] *Liu K., Yu C., Meng W., Wu W., Rishe N.* A Statistical Method for Estimating the Usefulness of Text Databases // IEEE TKDE, 2001.

[Luque, 2009] *Luque B., Lacasa L., Ballesteros F., Luque J.* Horizontal visibility graphs: Exact results for random time series // Phys. Review E, 2009. – P. 046103-1 – 046103-11.

[Lurie, 1983] *Lurie D., Valls J., Wagensberg J.* Thermodynamic approach to biomass distribution in ecological systems // Bull. Math. Biol, 1983. – **45**. – P. 869.

[*Manber, 1994*] *Manber U.* Finding similar files in a large file system // Proceedings of the 1994 USENIX Conference, January 1994. – P. 1-10.

[Meng, 2002] *Meng W., Yu C, Liu K.L.* Building Efficient and Effective Metasearch Engines // ACM Comput. Surv. 34, 1 (Mar. 2002), 48-89 pp.

[Milgram, 2009] *Milgram S.* The small world problem, Psychology Today, 1967. – **2**. – P. 60-67.

[Mostafa, 2013] Mostafa M.M. More than words: Social networks' text mining for consumer brand sentiments //

Expert Systems with Applications, 2013. – **40**. – № 10. – P. 4241–4251.

[Motter, 2002] *Motter A. E., de Moura A. P. S., Lai Y.-C., Dasgupta P.* Topology of the conceptual network of language // Phys. Rev., 2002. – **E 65**, 065102(R).

[Newman, 2003] Newman M.E.J. The structure and function of complex networks // SIAM Review, 2003. – **45**. – P. 167-256.

[Newman, 2006] *Newman M., Barabási A.-L., Watts D.J.* The Structure and Dynamics of Networks // Princeton and Oxford: Princeton University Press, 2006. – 624 p.

[Nowak, 1990] *Nowak A., Szamrej J., Latane B.* From private attitude to public opinion: A dynamic theory of social impact // Psychological Review, 1990. – 97. – P. 367-376.

[Nunez, 2012] *Nunez A.M., Lacasa L., Gomez J.P., Luque B.* Visibility algorithms: A short review // New Frontiers in Graph Theory, Y. G. Zhang, Ed. Intech Press, 2012. – Ch. 6. – P. 119 – 152.

[Ortuño, 2003] Ortuño M., Carpena P., Bernaola P., Muñoz E., Somoza A.M. Keyword detection in natural languages and DNA // Europhys. Lett., 2002. – **57**. – P. 759 – 764.

[Osipovs, 2009] *Osipovs P., Borisov A.* Practice of Web Data Mining Methods Application // J. Riga Technical University 40: 101-107 (2009) . – P. 11-18.

[Pastor-Satorras, 2001] *Pastor-Satorras R., Vespignani A.* Epidemic spreading in scale-free networks // Physics Review Letters, april 2001. – 86. – № 14.

[Price, 2001] *Price G., Sherman C., Sullivan D.* The Invisible Web: Uncovering Information Sources Search Engines Can't See // Information Today, Inc., 2001. – 439 p.

[Quinn, 2014] Quinn L.D., Endres A. B., Voigt T. B. Why not harvest existing invaders for bioethanol? // Biological Invasions, 2014. – **16**. – № 8. – P. 1559-1566.

[Sageman, 2004] *Sageman M.* Understanding Terror

Networks. – University of Pennsylvania Press, 2004.

[Salton, 1975] *Salton G, Wong A, Yang C.* A Vector Space Model for Automatic Indexing. // Communications of the ACM, 1975. – 18(11): 613-620.

[Salton, 1988] *Salton G., Buckley C.* Term-weighting approaches in automatic text retrieval. // Information Processing and Management, 1988. – **24**: 513-523.

[Shelton, 2003] *Shelton C., Koopman P., Nace W.* A framework for scalable analysis and design of system-wide graceful degradation in distributed embedded systems // Eighth IEEE International Workshop on Object-oriented Real-time Dependable Systems (WORDS 2003), Guadelajara, Mexico, Jan. 2003. – 8 p.

[Sigman, 1999] *Sigman M., Cecchi G.A.* Global Properties of the Wordnet Lexicon // Proc. Natl. Acad. Sci. USA, 1999, 1742.

[Sobkowicz, 2003] *Sobkowicz P.* Effect of leader's strategy on opinion formation in networked societies // Preprint Arxiv (on-line: http://arxiv.org/pdf/cond-mat/0311566)

[Stohl, 2007] *Stohl C., Stohl M.* Networks of Terror: Theoretical Assumptions and Pragmatic Consequences // Communication Theory. – 17 (2007). – P. 93-124.

[Stone, 2003] *Stone W.R.* Plagiarism, Duplicate Publication and Duplicate Submission: They Are All Wrong! // IEEE Antennas and Propagation, Aug. 2003. – 45. – № 4. – P. 47-49.

[Vapnik, 1998] *Vapnik V.N.* Statistical Learning Theory. – NY: John Wiley, 1998.

[Webb, 1995] *Webb J.N.* Hamilton's variational principle and ecological models // Ecoligical Modelling. 1995. – **80.** – P. 35.

[Yagunova, 2012] *Yagunova E., Lande D.* Dynamic Frequency Features as the Basis for the Structural Description of Diverse Linguistic Objects // CEUR Workshop Proceedings. Proceedings of the 14th All-

Russian Scientific Conference "Digital libraries: Advanced Methods and Technologies, Digital Collections" Pereslavl-Zalessky, Russia, October 15-18, 2012. – P. 150-159.

[Zhou, 2009] *Zhou Sh., Mondragon R.J.* The rich-club phenomenon in the Internet topology // Communications Letters, IEEE, March 2004. – 8, Issue 3. – P. 180-182.

[Zipf, 1949] *Zipf G.K.* Human Behavior and the Principle of Least Effort. – Cambridge, MA: Addison-Wesley Press, 1949. – 573 p.

[Анисимов, 2005] *Анисимов А.В., Тарануха В.Ю.* Возможности применения WordNet и других лингвистических онтологий в современных информационных системах // Автоматика-2005: Материалы 12-й международной конференции по автоматическому управлению. – Харьков: Изд. НТУ «ХПИ», 2005. – 3. – С. 56-57.

[Антонова, 2011] *Антонова А.Ю., Клышинский Э.С.* Об использовании мер сходства при анализе документации // Труды 13-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции – RCDL'2011», – 2011. – С. 134-138.

[Арнольд, 1990] *Арнольд В.А.* Теория катастроф. – М.: Наука, 1990. – 128 с.

[Астафьева, 1996] *Астафьева Н.М.* Вейвлет-анализ: основы теории и примеры применения // Успехи физических наук, 1996. – 166. – No 11. – P. 1145-1170.

[Ашура, 2006] *Аушра А.* Научная электронная библиотека как средство борьбы с плагиатом // Educational Technology & Society 9(3), 2006. – С. 270-276.

[Базак, 2003] *Базак Д.* Мета-поиск – лучший друг // [Электронный ресурс], 2003. Режим доступа: http://citforum.ru/internet/search/meta_poisk/

[Барсегян, 2003] *Барсегян А.А, Куприянов М.С., Степаненко В.В. Холод И.И.* Технологии анализа

данных: Data Mining, Visual Mining, Text Mining, Olap / 2-е изд., перераб. – СПб.: БХВ-Петербург, 2007. – 384 с.

[Брайчевский, 2005] *Брайчевский С.М., Ландэ Д.В.* Современные информационные потоки: актуальная проблематика // Научно-техническая информация, 2005. – Сер. 1. – Вып 11. – С. 21-33.

[Бреер, 2004] *Бреер В.В.* Стохастические модели социальных сетей // Управление большими системами. – 2009. – № 27. – С. 169-204.

[Вайдлих, 2005] *Вайдлих В.* Социодинамика: системный подход к математическому моделированию в социальных науках. – М.: Едиториал УРСС, 2005. – 480 с.

[Воронина, 2010] *Воронина И.Е., Пигалкова Е.А.* Создание базовой онтологии для российской системы права на основе онтологии LKIF-CORE // Вестник ВГУ, серия: Системный анализ и информационные технологии, 2010. – № 1. – С. 154-159.

[Гасфилд, 2003] *Гасфилд Д.* Строки, деревья и последовательности в алгоритмах: Информатика и вычислительная биология / пер. с англ. – СПб.: Невский Диалект; БХВ-Петербург, 2003. – 654 с.

[Головач, 2006] *Головач Ю., Пальчиков В.* Лис Микита і мережі мови // Журн. Фіз. Досл., 2006. – № 10. – С. 247-291.

[Горбулін, 2009] *Горбулін В.П., Додонов О.Г., Ланде Д.В.* Інформаційні операції та безпека суспільства: загрози, протидія, моделювання: монографія. – К.: Інтертехнологія, 2009. – 164 с.

[ГОСТ, 1990] Государственный стандарт Союза ССР. Система стандартов по информации, библиотечному и издательскому делу. Тезаурус информационно-поисковый многоязычный. Состав, структура и основные требования к построению ГОСТ 7.24-90, Москва, 1990.

[Григорьев, 2005] *Григорьев А.Н., Ландэ Д.В.* Адаптивный интерфейс уточнения запросов к системе контент-мониторинга InfoStream // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог-2005» (Звенигород, 1–6 июня, 2005 г.) – М.: Наука, 2005. – С. 109–111.

[Губанов, 2009] *Губанов Д.А., Новиков Д.А., Чхартишвили А.Г.* Модели информационного влияния и информационного управления в социальных сетях // Проблемы управления. – 2009. – № 5. – С. 28-35.

[Давыдов, 2008] *Давыдов А.А.* Системная социология. – М.: Издательство ЛКИ, 2008. – 192 с.

[Добров, 2009] *Добров Б.В., Соловьев В.Д., Лукашевич Н.В., Иванов В.В.* Онтологии и тезаурусы. Модели, инструменты, приложения. Бином, 2009. – 173 с.

[Додонов, 1990] *Додонов А.Г., Кузнецова М.Г., Горбачик Е.С.* Введение в теорию живучести вычислительных систем. – К: Наук. думка, 1990. – 184 с.

[Додонов, 2004] *Додонов А.Г., Флейтман Д.В.* К вопросу безопасности информационных систем // Збірник наукових праць «Інформаційні технології та безпека» – 2004. – Вип. 6. Київ. – С. 26-29.

[Додонов, 2006] *Додонов А.Г., Ландэ Д.В.* Организация сети информационных прокси-серверов // Реєстрація, зберігання і обробка даних, 2006. – **8. –** № 3. – С. 24-31.

[Додонов, 2009] *Додонов О.Г., Ланде Д.В., Путятін В.Г.* Інформаційні потоки в глобальних комп'ютерних мережах. – К.: Наукова думка, 2009. – 295 с.

[Додонов, 2011] *Додонов А.Г., Ландэ Д.В.* Живучесть информационных систем. – К.: Наук. думка, 2011. – 256 с.

[Додонов, 2013] *Додонов А.Г., Ландэ Д.В., Прищепа В.В., Путятин В.Г.* Конкурентная разведка в компьютерных сетях.– К.: ИПРИ НАН Украины, 2013. – 248 с.

[Додонов, 2013-1] *Додонов А.Г., Ландэ Д.В., Путятин В.Г., Жигало В.В.* Архітектура системи моніторингу, адаптивного агрегування та узагальнення інформації // Реєстрація, зберігання і обробка даних. - 2013, - Т.15. - №4. – С. 32-40.

[ДСТУ, 2001] ДСТУ 4032-2001 «Одномовний тезаурус. Методика розроблення».

[Дубичинский, 2008] *Дубичинский В.В.* Лексикография русского языка: учеб. пособие. – М.: Наука, Флинта, 2008. – 432 с.

[Зеленков, 2007] *Зеленков Ю.Г., Сегалович И.В.* Сравнительный анализ методов определения нечетких дубликатов для Web-документов // Труды 9-ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2007, Переславль, Россия, 2007. –**1**. С. 166-174.

[Иванов, 2002] *Иванов С.А.* Стохастические фракталы в информатике // Научно-техническая информация, 2002. – Сер. 2. – Вып 8. – С. 7-18.

[Капица, 1977] *Капица С.П., Курдюмов С.П., Малинецкий Г.Г.* Синергетика и прогнозы будущего. – М.: Наука, 1997. – 288 с.

[Коваленко, 2013] Коваленко Е.А. Введение в теорию информационного пространства организации // «Экономика и современный менеджмент: теория и практика»: сборник статей по материалам XXIX международной научно-практической конференции (11 сентября 2013 г.)

[Кононов, 2003] *Кононов Д.А., Кульба В.В., Шубин А.Н.* Базисные понятия моделирования информационного управления в социальных системах // Труды международной научно-практической конференции «Теория активных систем». – М.: Институт проблем управления им. В.А. Трапезникова РАН, 2003. – **2**. – С. 125-129.

[Кормен, 2006] *Кормен Т.Х., Лейзерсон Ч., Ривест Р., Штайн К.* Алгоритмы: построение и анализ. – 2-е изд. – М.: «Вильямс», 2006. – 1296 с.

[Кульба, 2004] *Кульба В.В., Кононов Д.А., Косяченко С.А., Шубин А.Н.* Методы формирования сценариев развития социально-экономических систем. – М.: СИНТЕГ, 2004. – 296 с.

[Ланде, 1999] *Ланде Д.В., Сороко В.М.* Створення основ функціонального класифікатора з питань державної служби в Україні // Вісник державної служби України, 1999. – № 4. – С. 83-88.

[Ланде, 2012] *Ланде Д.В.* Методи оцінки рівня дискримінантної сили слів у текстах з правової тематики // Правова інформатика, 2012. – № 3 (35). – С. 5-9.

[Ланде, 2013] *Ланде Д.В.* Тренди відображення інформаційних операцій в інформаційному просторі // Інформація і право, 2013. – N 1 (7). – С. 82-88.

[Ланде, 2014] *Ланде Д.В.* Елементи комп'ютерної лінгвістики в правовій інформатиці. – К.: НДІІП НАПрН України, 2014. – 168 с.

[Ландэ, 2006] *Ландэ Д.В.* Основы интеграции информационных потоков: – К.: Инжиниринг, 2006. – 240 с.

[Ландэ, 2006-1] *Ландэ Д.В., Дармохвал А.Т., Морозов А.Ю.* Подход к выявлению дублирования сообщений в новостных информационных потоках // Труды 8-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции – RCDL'2006», 2006. – С. 115-119.

[Ландэ, 2007] *Ландэ Д.В., Снарский А.А., Брайчевский С.М., Дармохвал А.Т.* Моделирование динамики новостных текстовых потоков // Интернет-математика 2007: Сборник работ участников конкурса. –Екатеринбург: Изд-во Урал. ун-та, 2007. – С. 98-107.

[Ландэ, 2007-1] *Ландэ Д.В.* Модель диффузии информации // Информационные технологии и безопасность. Менеджмент информационной безопасности. Сборник научных трудов Института проблем регистрации информации. – Вып. 10. – 2007. – С. 51-67.

[Ландэ, 2008] *Ландэ Д.В., Брайчевский С.М., Дармохвал А.Т., Морозов А.Ю.* Ранжирование источников информации в системе мониторинга новостей InfoStream // Труды 10-й Всероссийской научной конференции "Электронные библиотеки: перспективные методы и технологии, электронные коллекции" – RCDL'2008, Дубна, Россия, 2008. – С 213-219.

[Ландэ, 2008-1] *Ландэ Д.В., Брайчевский С.М., Дармохвал А.Т., Морозов А.Ю.* Веб-пространство и материалы информационных агентств // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции "Диалог". Вып. 7 (14). – М.: Изд-во РГГУ, 2008. – С. 303-305.

[Ландэ, 2009] *Ландэ Д.В., Снарский А.А., Безсуднов И.В.* Интернетика: Навигация в сложных сетях: модели и алгоритмы. – М.: Либроком (Editorial URSS), 2009. – 264 с.

[Ландэ, 2010] *Ландэ Д.В., Снарский А.А., Жигало В.В.* Метапоиск доступных научно-технических документов в Интернет // Труды 12-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2010, Казань, Россия, 2010. – С 321-325.

[Ландэ, 2012] *Ланде Д.В., Фурашев В.М.* Основи інформаційного і соціально-правового моделювання: монографія. – К.: ТОВ "ПанТот", 2012. – 144 с.

[Ландэ, 2013] *Ландэ Д.В.* Метод визуализации зон нестабильности в рядах измерений // Информационные

технологии и безопасность. Оценка состояния: Материалы международной научной конференции ИТБ-2013. – К.: ИПРИ НАН Украины, 2013. – С. 105-113.
[Ландэ, 2014], *Ландэ Д.В., Снарский А.А., Путятин В.Г.* Построение терминологической сети предметной области // – Реєстрація, зберігання і обробка даних, - 2014. - Т.16. №2. – С. 114-121.

[Левич, 1980] *Левич А.П.* Структура экологических сообществ. – М.: Изд-во Моск. ун-та, 1980. – 181 с.

[Манойло, 2003] *Манойло А.В.* Государственная информационная политика в особых условиях / А.В. Манойло. — Монография. — МИФИ, 2003. — 388 с.

[Массон, 2004] *Массон Г.В.* Взаимосвязь системы личностных терминальных ценностей и типов межличностных отношений: Дис. … канд. психол. наук: 19.00.01: Красноярск, 2004. – 146 с. РГБ ОД, 61:05-19/11.

[Морозов, 1915] *Морозов Н.А.* Лингвистические спектры: средство для отличения плагиатов от истинных произведений того или иного известного автора. Стилеметрический этюд. // Известия отд. русского языка и словестности Имп. Акад. наук, Т. XX, кн. 4, 1915.

[Нежданов, 2009] *Нежданов И.* Технологии разведки для бизнеса. – М.: Ось-89, 2009. – 400 с.

[Нейл, 2005] *Нейл К., Шанмагантан Г.* Веб-инструмент для выявления плагиата // Открытые системы, 2005. – № 1. – С. 40-44.

[Никконен, 2007] *Никконен А.Ю.* Устранение избыточности и дублирования сюжетов новостных сообщений // Интернет-Математика. Сборник работ участников конкурса. – Екатеринбург: Изд-во Урал. Ун-та, 2007. – С. 157-167.

[Плотинский, 2006] *Плотинский Ю.М.* Модели социальных процессов. – Изд. 2-е. – М.: Логос, 2001. – 296 с.

[Потеев, 1999] *Потеев М.И.* Концепции современного естествознания. – СПб.:, Издательство «Питер», 1999. – 352 с.

[Потемкин, 2008] *Потемкин С.Б., Кедрова Г.Е.* Выравнивание неразмеченного корпуса параллельных текстов // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2008». – Вып. 7 (14). – М.: РГГУ, 2008. – С. 431-436.

[Пригожин, 1986] *Пригожин И., Стенгерс И.* Порядок из хаоса. – М., 1986. – 432 с.

[Приц, 1974] *Приц А.К.* Принцип стационарных состояний открытых систем и динамика популяций. Калининград, 1974. – 123 с.

[Райншке, 1979] *Райншке К.* Модели надежности и чувствительности систем. – М.: Мир, 1979. – 454 с.

[Райншке, 1998] *Райншке К., Ушаков И.* Оценка надежности систем с использованием графов. – М. : Радио и связь, 1988. – 208 с.

[Самойленко, 2001] *Самойленко А.П., Дюк В.А.* Data Mining: Учебный курс. CD. – СПб: Питер.: 2001. – 368 с.

[Свирежев, 1991] *Свирежев Ю.М.* Феноменологическая термодинамика взаимодействующих популяций // Журн. общ. биологии. 1991. – **52**. – № 6. – С. 840.

[Снарский, 2009] *Снарский А.А.* Метод выявления неявных связей объектов / А.А. Снарский, Д.В. Ландэ, М.И. Женировский // Труды 11-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2009, Петрозаводск, 2009. – С. 46-49.

[Соловьев, 2006] *Соловьев В.Д.* Онтологии и тезаурусы / В.Д. Соловьев, Б.В. Добров, В.В. Иванов, Н.В.

Лукашевич. – Казань: Казанский государственный университет, 2006. – 157 с.

[Сорока, 1998] *Сорока М., Танатар Н.В.* Використання методу контент-аналізу при створенні автоматизованих інформаційних систем // Наук. пр. НБУВ. – 1998. – Вып. 1. – С. 318-323.

[Федер, 1991] *Федер Э.* Фракталы. – М.: Мир, 1991. – 254 с.

[Фурсова, 2003] *Фурсова П.В., Левич А.П., Алексеев В.Л.* Экстремальные принципы в математической биологии // Успехи современной биологии, 2003. – Т. 123. – № 2. – С. 115-117.

[Чубукова, 2006] *Чубукова И.А.* Data Mining: учебное пособие. – М.: Интернет-университет информационных технологий ИНТУИТ.ру, 2006. – 384 с.

[Ханин, 1982] *Ханин М.А.* Энергетика и критерии оптимальности онтогенетических процессов. Математическая биология развития. – М.: Наука, 1982. – 177 с.

[Хорошевский, 2013] *Хорошевский В.Ф.* Семантические технологи: ожидания и тренды // Открытые Семантические технологии проектирования интеллектуальных систем – Open Semantic Tech-nologies for Intelligent Systems (OSTIS-2012): материалы II Междунар. научн.-техн. конф. (Минск, 16-18 февраля 2012 г.). – Минск: БГУИР, 2012. – С. 143-158.

[Чертов, 2009] *Четов О.Р.* Поліноми Кунченка для розпізнавання образів // Вісник НТУУ «КПІ» Інформатика, управління та обчислювальна техніка, 2009. – № 50. – С. 105-110.

[Чхартишвили, 2004] *Чхартишвили А.Г.* Теоретико-игровые модели информационного управления. – М.: ЗАО «ПМСОФТ», 2004. – 227 с.

[Шарапов, 2011] *Шарапов Р.В.* Анализ подходов к обнаружению заимствованных текстов //

Фундаментальные исследования. – 2011. – № 3 – С. 47-49.

[Щерба, 1936] *Щерба Л.В.* Передмова до російсько-французського словника / Русско-французский словарь / Сост. Л. В. Щерба, М. И. Матусевич, М. Ф. Дусс. Под общ. рук. и ред. Л. В. Щербы. – М., 1936. 11 с. без пагинации – 491 с.

[Широков, 1998] *Широков В.А.* Інформаційна теорія лексикографічних систем. – К.: Довіра, 1998. – 331 с.

[Широков, 2005] *Широков В.А.* Елементи лексикографії.– К.: Довіра, 2005. – 304 с.

[Широков, 2005а] *ШироковВ.А., Бугаков О.В., Грязнухіна Т.О.* Корпусна Лінгвістика. – К.: Довіра, 2005. – 471 с.

[Широков, 2011] *Широков В.А.* Комп'ютерна лексикографія. – К.: Наукова думка, 2011. – 352 с.

[*Шрейдер, 1971*] *Шрейдер Ю.А.* Равенство, сходство, порядок. – М.: Наука, 1971. – 256 с.

[Ягунова, 2010] Ягунова Е.В. Эксперимент и вычисления в анализе ключевых слов художественного текста // Сборник научных трудов кафедры иностранных языков и философии ПНЦ УрО РАН. Философия языка. Лингвистика. Лингводидактика. – Пермь, 2010. – Вып. 1. – С. 85-91.

ГЛОССАРИЙ

Автоматическое реферирование (Summarization) – автоматическое формирование краткого изложения исходного текстового материала путем: 1) выделения фрагментов информационного наполнения и последующего их соединения; 2) методом генерации текста на основании выявления знаний из оригинала.

Агрегатор новостей (News aggregator) – интернет-сервис, обеспечивающий доступ пользователей или приложений к актуальной информации из многих новостных источников. Агрегатор новостей собирает RSS потоки и предоставляет их читателю на одной странице, таким образом, можно настроить новостную ленту или внешний вид новостной страницы сайта под себя.

Агрегирование информации (Information aggregation) – концентрирование отдельных информационных потоков в единый сводный агрегат, что дает возможность получить общую картину ситуации.

Адаптация (Adaptation) – приспособление системы к условиям изменяющейся внешней среды.

Адаптивность (Adaptivity) – способность системы модифицировать себя или свое окружение при изменении условий функционирования для компенсации потери эффективности функционирования; способность системы приспосабливаться к различным условиям окружающей среды.

Адекватность аналитической модели (Adequacy of a model) – степень соответствия модели реальному объекту или процессу, полнота и точность описания ею предмета исследования.

Адекватность информации (Information adequacy) – уровень соответствия создаваемого с помощью

полученной информации образа реальному объекту, процессу, явлению и т.п.

Алгебраическая модель информационного поиска (Algebraic model of information retrieval) – модель информационного поиска, в которой документы и запросы описываются в виде векторов в многомерном пространстве. Каркасом для таких моделей выступают алгебраические методы.

Алгоритм поиска (Search algorithm) – описание заранее предопределенной последовательности действий поисковой системы для отбора информации по запросу пользователя.

Анализ данных (Data analysis) – действия с данными, направленные на извлечение из них содержащейся информации об исследуемом объекте и на получение новых данных на основании имеющихся.

Анализ текста (Text analysis) – процесс извлечения информации из текстовых данных на основе обнаружения в них закономерностей.

Аналитическая деятельность (Analytical work) – обусловленная поставленными целями и задачами совокупность технологий интеллектуальной деятельности, обеспечивающих эффективную обработку информации, установление ее полноты, достоверности и иных свойств, выявление проблем и тенденций развития процессов, а также подготовку управленческих решений.

Аналитическая модель (Analytical model) – математическая модель, характеризующая функциональные зависимости результатов (выходов) от параметров (входов).

Аналитическая справка (Background) – результат работы информационно-аналитической службы по определенным вопросам социально-экономического, политического характера, проблемам безопасности, перспективам развития предприятия, фирмы в определенном сроке времени и пространстве.

Аналитический запрос (Analytical request) – запрос к базе или хранилищу данных, целью которого является получение не просто нужного отчета, но и полезных выводов и знаний на основе содержащейся в нем информации.

Анафора (Anaphora) – под анафорой в синтаксисе понимается такой способ построения текста, при котором отдельные элементы или все содержание одного, обычно предшествующего, предложения воспроизводится в другом, обычно последующем предложении. Средства такого воспроизведения могут быть различными. В простейших формах речи это чаще всего – лексический повтор; использование языковых выражений, которые могут быть проинтерпретированы лишь с учетом другого, как правило, предшествующего, фрагмента текста.

Анафорические связи (Anaphoric links) – отношения между частями текста (между словами, словосочетаниями, высказываниями), при которых в смысл одного слова (словосочетания, высказывания) входит отсылка к другому слову (словосочетанию, высказыванию).

Антонимы (Antonyms) – слова одной части речи, различные по звучанию и написанию, имеющие прямо противоположные лексические значения.

Апплет (Applet) – программа, встраивается в веб-сайт; передается клиенту из веб-сервера вместе с документами в виде добавления, обеспечивающего их представление пользователям.

Аскриптор (Askriptor) – лексическая единица, подлежащая замене на дескриптор в поисковых образах документов (поисковых образах запросов) при поиске и обработке информации.

Аттрактор (Attractor) – компактное подмножество фазового пространства динамической системы, все траектории из некоторой окрестности которого стремятся к нему при времени, стремящемся к

бесконечности. Наиболее простыми вариантами аттрактора являются притягивающая неподвижная точка и периодическая траектория.

База данных поисковой системы (Search engine database) – место хранения основных параметров (индексов) каждого известного данной системе документа; пополняется поисковым роботом во время периодических обходов веб-пространства.

База знаний (Knowledge base) – совокупность фактов и правил, допускающих автоматические выводы и обработку информации.

Банк данных (Data bank) – автоматизированная информационная система централизованного хранения и коллективного использования данных. В состав банка данных входят одна или несколько баз данных, справочник баз данных, СУБД, а также библиотеки запросов и прикладных программ.

Безмасштабная сеть (Scale-free network) – это сеть, в которой степени вершин распределены по степенному закону или закону, приближающемуся к степенному в асимптотике. Это означает, что доля узлов в сети имеющих k связей составляет $P(k) \sim ck^{-\gamma}$ для больших значений k. Здесь c – нормализующий параметр, а γ – индивидуальная характеристика сети.

Бифуркация (точка бифуркации) (Bifurcation (bifurcation point)) – понятие, описывающее неравновесное состояние системы, из которого равновероятным является переход к одному из возможных сценариев ее изменения.

Блог (Blog) – сетевой дневник одного или нескольких авторов, состоящий из записей в обратном хронологическом порядке. С помощью сервиса блогов можно создать онлайн-дневник, читать и комментировать дневники других пользователей, принимать участие в сообществах и создавать свои сообщества.

Валидация модели (Validation of the model) – проверка правильности работы модели, построенной на основе машинного обучения, а также удостоверение о ее соответствии требованиям решаемой задачи. Валидация проводится на независимом (т.е. не использовавшемся для обучения и тестирования) валидационном множестве после обучения и тестирования модели.

Валидность (Validity) – обоснованность и адекватность исследовательских инструментов. Различают внутреннюю и внешнюю валидность.

Веб–аналитика (Web analytic) – измерение, сбор, анализ, представление и интерпретация информации о посетителях веб-сайтов в целях их улучшения и оптимизации. Задача веб-аналитики – мониторинг работы веб-сайтов, на основании которого определяется веб-аудитория и изучается поведение веб-посетителей для принятия решений по развитию и расширению функциональных возможностей веб-ресурса.

Веб-документ (веб-страница) (Webpage; Web document) - составная часть веб-сайта, это электронный документ, который может содержать текст, изображения, Java апплеты и другие элементы. Веб-документ может быть статическим или динамически сгенерированным.

Веб-сайт (Website) – набор веб-документов, составляющих некоторое единство, как правило, размещенных на одном и том же сервере, имеющих одно и то же доменное имя и связанных между собой перекрестными ссылками.

Вейвлет (Wavelet) – класс математических функций, позволяющих анализировать различные частотные компоненты данных. Основное свойство вейвлетов – локализация во времени и по частоте, что дает возможность строить на основе одного и того же вейвлета семейство функций посредством его сдвигов и растяжений по оси времени.

Вейвлет-анализ (Wavelet-analisis) – анализ данных с использованием вейвлет-преобразований.

Вейвлет-преобразование (Wavelet transform) – преобразование, функции, которые рассматривают ее в терминах колебаний, локализованных по времени и частоте. Главная идея вейвлет-преобразования заключается в том, что нестационарный временной ряд делится на отдельные промежутки (так называемые «окна наблюдения»), и на каждом из которых выполняется вычисление скалярного произведения исследуемых данных с различными сдвигами некоторого вейвлета на разных масштабах. Вейвлет-преобразование обычно разделяют на дискретное (DWT) и непрерывное (CWT) вейвлет-преобразование.

Верификация (Verification) – проверка; способ обоснования (подтверждения) каких-либо теоретических положений путем их сопоставления с опытными (эмпирическими) данными.

Вес (Weight) – величина (коэффициент), которая характеризует значимость объекта среди подобных.

Весовой коэффициент (Weight coefficient) – в компьютерной лингвистике – коэффициент, приписываемый лексической единице в документе. Может зависеть от расположения лексической единицы в документе, абзаце, предложении. Непосредственно зависит от смысла лексической единицы, частоты встречаемости.

Визуализация (Visualization) – комплекс методов представления результатов анализа данных в наиболее удобной для восприятия и интерпретации форме.

Визуализация данных (Data visualization) – проектирование и генерация изображений на устройствах отображения на основе исходных цифровых данных, а также правил и алгоритмов их преобразования.

Внешняя среда (Environment) – объекты, не принадлежащие рассматриваемому объекту, но влияющие на него.

Временной ряд (Time series) – данные, последовательно измеренные через некоторые (зачастую равные) промежутки времени.

Вторичная информация (Secondary information) – информация, полученная в результате обработки первичной информации.

Вторичные данные (Secondary data) – данные, которые являются результатом определенных вычислений, примененных к первичным данным.

Вторичный документ (Secondary document) – документ, полученный в результате аналитико-синтетической и логической переработки сведений или данных, содержащихся в первичных документах.

Входной документ (Input document) – документ, составленный по определенной форме и содержащий данные, предназначенные для ввода в память компьютера.

Входной поток (Input stream) – последовательность документов или данных, поступающих для ввода в автоматизированную информационную систему.

Входной файл (Input file) – файл, содержащий входные данные.

Входные данные (Input data) – данные, вводимые в вычислительную систему через устройства ввода для обработки или хранения.

Выборка (Sample) – множество объектов, отобранное из большого количества объектов генеральной совокупности, отражающих этих качества объектов.

Выборка данных (Data sampling) – процесс поиска и считывания данных из файла, группы файлов или базы данных.

Выходной документ (Output document) – документ, являющийся носителем результатов обработки данных или формируемый автоматизированной системой и выданный системными средствами вывода.

Гиперсвязь, гиперссылка (Hyperlink) – связь между отдельными компонентами информации. Применяется для организации ссылок, сделанных внутри одного объекта или от одного объекта на другой. Гиперссылки связывают страницы веб-сайта, тем самым они является основой его структуры.

Гипертекст (Hypertext) – текстовый документ, содержащий гиперссылки на другие документы (или имеющий внутренние связи). Представляет собой специальным образом размеченную текстовую информацию.

Глобальные сети (Wide area network) – телекоммуникационные структуры, объединяющие локальные информационные сети, имеющие общий протокол связи, методы подключения и протоколы обмена данными.

Глубинный анализ данных (Data mining) – технология анализа данных в БД или хранилищах данных, основанная на статистических методах и служащая для выявления заранее неизвестных закономерностей, а также для поддержки принятия стратегически важных решений.

Глубинный анализ текста (Text mining) – процесс извлечения информации из текстовых данных на основе обнаружения в них закономерностей. Этот анализ, как правило, включает этапы структурирования исходного текста, поиска закономерностей в данных, а также оценивания и интерпретации результатов.

Графическая информация (Graphic information) – сведения или данные, представленные в виде схем, эскизов, изображений, графиков, диаграмм, символов.

Графический файл (Graphic file) – файл, содержащий графическое изображение.

Группирование (Grouping) – объединение нескольких объектов в один общий.

Дактилограмма (Dactylogram) – подстрока документа фиксированной длины.

Двоичный поиск (Binary search) – алгоритм поиска объекта по заданному признаку во множестве объектов, упорядоченных по тому же самому признаку. Двоичный поиск заключается в том, что на каждом шаге множество объектов делится на две равные части и в работе остается та часть множества, где находится искомый объект.

Дезинформация (Disinformation) – передаваемые кому-либо и в любой форме данные, сведения, сообщения и т.п., неверно отражающие объекты описания.

Дерево (Tree) – в теории графов – связный граф с одной вершиной (корневой вершиной), в которую нет входящих ребер, а в каждую другую вершину входит только одно ребро.

Дескриптор (Descriptor) – лексическая единица (слово, словосочетание, код) информационно-поискового языка, служащая для выражения основного смыслового содержания документа (текста).

Дескрипторный словарь (Descriptor dictionary) – словарь информационно-поискового языка, в котором приведены в общем алфавитном ряду дескрипторы и их синонимы без указания других отношений лексических единиц. Дескрипторный словарь является упрошенным вариантом информационно-поискового тезауруса.

Дестабилизирующий фактор (Destabilizing factor) – явление или событие, следствием наступления которого может быть нарушение конфиденциальности, целостности и/или доступности информационных

ресурсов, нарушению работоспособности системы или ее элементов.

Детерминированный (динамический) хаос (Deterministic (dynamic) chaos) – явление, при котором поведение нелинейной системы выглядит случайным, несмотря на то, что она определяется детерминистическими законами. Причиной появления детерминированного хаоса является неустойчивость системы относительно начальных условий и параметров: изменение начального условия приводит к существенным изменениям динамики системы.

Диссипация (Dissipation) – процесс рассеяния чего-либо, перехода элементов системы из связного состояния, в состояние, характеризующееся наличием слабых или полной потерей связей и утратой упорядоченности системы. В диссипативных системах под действием внешних потоков энергии или информации происходят процессы самопроизвольного образования временных когерентных систем (самоупорядочения).

Диффузия информации (Diffusion of information) - взаимное проникновение различных информационных сообщений по аналогии с тем, как в физике происходит взаимное проникновение частиц соприкасающихся веществ. Процессы диффузии информации, как и процессы диффузии в физике, достаточно точно моделируются с помощью метода клеточных автоматов.

Документальная информационно-поисковая система (Documentary information retrieval system) – информационно-поисковая система (ИПС) для поиска документов, содержащих необходимую информацию. Поисковый массив документальной ИПС состоит из поисковых образов документов или из самих документов.

Документальная информация (Documentary information) – информация, основанная на документах, на фактах; информация, закреплённая посредством

какой-либо знаковой системы на материальном носителе.

Документальный информационный поток (Documentary information flow) – множество первичных и вторичных документов, а также источников информации, целенаправленно передающихся по информационным каналам от отправителя к потребителю.

Документальный источник (Documentary source) - источник информации о фактах, событиях, явлениях реального мира и мыслительной деятельности человека, закрепленных различными способами на специальном носителе.

Единица выборки (Sampling unit) – элемент генеральной совокупности, выступающий в качестве единицы счета при различных процедурах формирования выборки.

Живучесть информации (Information survivability) – способность информации сохранять свое качество с течением времени.

Жизненный цикл информации (Information life cycle) – период времени с момента создания информационного продукта (документа или ресурса) до его устаревания. Последнее может быть связано с потерей актуальности содержащихся в нем сведений, появлением новых и более точных данных и т.д.

Знание (Knowledge) – 1) информация, которая может быть полезна, понятна и доступна индивидууму (группе индивидуумов) при решении им задач; 2) совокупность информации о различных областях реальности, когнитивная основа человеческой деятельности.

Знания декларативные (Deklarative knowledge) – знания, которые записаны в памяти интеллектуальной системы так, что они непосредственно доступны для использования после обращения к соответствующему полю памяти.

Знания корпоративные (Corporate knowledge) – служебная информация, необходимая для поддержки на высоком уровне основных технологических процессов корпорации, а также для быстрого реагирования на динамику измерений.

Знания о предметной области (Domain knowledge) – совокупность сведений о предметной области, хранящихся в базе знаний интеллектуальной системы.

Извлечение знаний (Knowledge extracting) – процесс получения из данных знаний в виде зависимостей, правил, моделей. Этапы: консолидация, очистка, трансформация, моделирование и интерпретация полученных результатов.

Извлечение информации (Information extraction) – разновидность информационного поиска, при которой из электронных документов выделяется некая структурированная информация, т.е. категоризированные, семантически значимые данные по какой-либо проблеме или вопросу.

Индекс информационно-поисковой системы (Information retrieval system index) – определенным образом организованная совокупность данных, где содержатся поисковые образы всех документов базы данных. Является основной составляющей архитектуры информационно-поисковой системы.

Индексирование (Indexing) – 1) процесс описания содержания документов и запросов в терминах информационно-поискового языка; 2) процесс выражения главного предмета или темы документа на информационно-поисковом языке.

Инсайдер (Insider) – член какой-либо группы людей, имеющей доступ к информации, недоступной широкой публике.

Инсайдерская информация (Insider information) — существенная, публично не раскрытая служебная информация компании, которая в случае её раскрытия способна повлиять на рыночную позицию компании. В

более широком смысле — любая информация, известная неопределенному кругу лиц, близких к её источнику.

Интеграция данных (Data integration) – процесс комбинирования нескольких наборов данных для их последующего совместного использования и анализа в целях поддержки управления информацией в пределах компании.

Интерфейс (Interface) – система или подсистема, реализующая посреднические функции между различными системами; решает задачу взаимнооднозначного отображения множеств специфических сигналов, необходимых для коммуникации между системами.

Информант (Informant) – лицо, от которого получена информация; в лингвистике – носитель языка, от которого получают сведения, особенно о языках, не имеющих письменности, устойчивой литературной традиции.

Информационная война (Information war) – целенаправленные действия, предпринятые для достижения информационного превосходства путем нанесения ущерба информации, информационным процессам и информационным системам противника при одновременной защите собственной информации, информационных процессов и информационных систем.

Информационная живучесть (Information survivability) – способность системы поддерживать доступность, целостность и конфиденциальность информации на уровне, позволяющем выполнять с заданным качеством цель функционирования системы, независимо от внешних и внутренних неблагоприятных воздействий и нарушений в использовании информационных ресурсов.

Информационная операция (Information operation) – комплекс согласованных и взаимосвязанных мероприятий по манипулированию информацией,

осуществляемых по общему плану с целью достижения и удержан превосходства через воздействия на информационные процессы в системах противника. Использование сложной совокупности согласованных, скоординированных и взаимоувязанных по целям, задачам, месту и времени, объектам и процедурам видов, форм, способов и приемов информационного воздействия.

Информационная поддержка (Information Support) – процесс информационного обеспечения, ориентированный на пользователей информации, занятых управлением сложными объектами. Информационная поддержка используется при подготовке и реализации управленческих решений.

Информационная сеть (Information network) – совокупность взаимодействующих друг с другом информационных систем.

Информационная среда (Information environment) – совокупность технических и программных средств хранения, обработки и передачи информации, а также социально-экономических и культурных условий реализации процессов информатизации.

Информационная сфера (Information sphere) – сфера деятельности субъектов, связанная с созданием, преобразованием и потреблением информации.

Информационное агентство (News agency) – специализированное информационное предприятие (организация, служба, центр), обслуживающее средства массовой информации. Его основная функция – снабжать оперативной информацией редакции газет, журналов, телевидения, радиовещания, а также другие учреждения, организации, частных лиц, являющихся подписчиками на их продукцию. Функционирование информационных агентств ориентировано на сбор новостей.

Информационное воздействие (Information influence) – воздействие на массовое сознание

аналогичное тому, как психологическое воздействие влияет на сознание индивидуальное. Возбуждение (торможение) в управляемой системе таких процессов, которые стимулируют желательный для воздействующей стороны выбор. Этот способ воздействия на противника не предполагает прямого выведения из строя части элементов его системы, но представляет собой передачу противнику такой информации, которая натолкнет его на выбор необходимого для воздействующей стороны решения.

Информационное сообщение (Informational message) – 1) сообщение, предоставляющее пользователю некоторую информацию о системе, по сравнению с сообщениями о завершении, означающими успешное выполнение некоторой операции, а также аварийными и диагностическими сообщениями, описывающими сбой или невозможность выполнения запроса; 2) группа простейших элементов информации, имеющих внутреннюю взаимосвязь; 3) сводка новостей из жизни компаний, госведомств, политических, общественных и иных организаций.

Информационное хранилище (Data warehouse) – предметно-ориентированное, привязанное ко времени и постоянное хранилище данных для поддержки процесса принятия управленческих решений.

Информационно-аналитическая деятельность (Information and analytical deyatelnostlnust) – область человеческой деятельности, призванная обеспечить информационные потребности общества с помощью аналитических и информационных технологий, за счет переработки исходной информации и получения качественно нового знания.

Информационно-поисковая система (ИПС, Information Retrieval System, IRS) – система, предназначенная для обеспечения поиска и отображения документов, представленных в базах данных. Совокупность методов и средств, обеспечивающих осуществление информационного поиска.

Информационно-поисковый тезаурус (Information retrieval thesaurus) – нормативный словарь дескрипторного информационно-поискового языка с зафиксированными парадигматическими отношениями лексических единиц, указывающими общность или противопоставление значений и использования лексических единиц.

Информационные поводы (Informational reasons) – некие события (физически или виртуально происходящие), информация о которых может быть интересна средствам массовой информации.

Информационные ресурсы (Information resources) – отдельные документы и отдельные массивы документов, документы и массивы документов в информационных системах.

Информационный барьер (Information barrier) – препятствие, мешающее оптимальному протеканию информационных процессов. Различают: объективные информационные барьеры, возникающие и существующие независимо от человека; субъективные информационные барьеры, создаваемые источником информации; и субъективные информационные барьеры, возникающие за счет приемника информации.

Информационный массив (Informative array) – совокупность зафиксированной информации, предназначенная для хранения и использования и рассматриваемая как единое целое.

Информационный поток (Information stream) – совокупность сведений, циркулирующих как в информационной системе (ИС), та и между ИС и внешней средой.

Информационный элемент (Information element) – единица информации, подлежащей обработке и передаче пользователям системы или предназначенная для обеспечения ее работы.

Искажение информации (Distortion of information) – случайная несанкционированная модификация

информации при ее обработке техническими средствами в результате внешних воздействий (помех), сбоев в работе аппаратуры или неумелых действий обслуживающего персонала.

Источник информации (Information source) – объект, идентифицирующий происхождение информации.

Итерация (Iteration) – метод решения задачи последовательным приближением к правильному результату. Итерация основана на повторении последовательности операций, при котором на каждом шаге повторения используется результат предыдущего шага.

Капча (Captcha, Completely Automated Public Turing test to tell Computers and Humans Apart) — полностью автоматизированный публичный тест Тьюринга – компьютерный тест, используемый для того, чтобы определить, кем является пользователь системы.

Классификация (Classification) – система распределения объектов по классам в соответствии сопределенным признаком (основание классификации). Объекты необходимо классифицировать для выявления общих свойств информационного объекта, который определяется информационными параметрами.

Кластеризация (Clustering) – метод анализа данных, основанный на группировании записей в соответствии с их расположением в пространстве многомерных атрибутов. Объекты внутри кластера должны быть похожими друг на друга и отличаться от объектов, вошедших в другие кластеры.

Клеточный автомат (Cellular automata) – набор клеток, образующих некоторую периодическую решетку с заданными правилами перехода, которые определяют состояние клетки в следующий момент времени через состояние клеток, находящимися от нее на расстоянии не больше некоторого заданного, в текущий момент времени. Как правило, рассматриваются автоматы, где

состояние определяется самой клеткой и ближайшими соседями.

Ключевое слово (Keyword) – лексическая единица, отдельное слово или словосочетание, которое используется при индексировании документов и поиске в информационно-поисковых системах.

Компактификация (Compactication) – операция, которая преобразует произвольные топологические пространства в компактные.

Контекст (Context) – часть текста, позволяющая определить значение какого-либо слова или фразы; законченный отрывок текста, общий смысл которого позволяет уточнить значение отдельных входящих в него слов, предложений.

Контекстный анализ объектов (Contextual analysis of objects) – поиск в текстовом массиве всех связей указанного объекта, а также всех объектов, связанных с исходным, с возможностью получения исходных документов, содержащих описания обнаруженных объектов.

Контент (Content) – содержательная часть информационных ресурсов. Существенными параметрами контента является его объем, актуальность и релевантность.

Контент-анализ (Content analysis) – метод получения выводов путем анализа содержания текстовой информации. Реализуется как систематическая обработка, оценка и интерпретация формы и содержания информационного источника. Результаты могут использоваться в технологиях интеллектуального анализа текстовых данных Text Mining.

Контент-мониторинг (Content monitoring) – систематическое, непрерывное во времени сканирование и контент-анализ информационных ресурсов.

Корпус лингвистический (Corps linguistic) – совокупность текстов, собранных в соответствии с определенными принципами, размеченных по определенному стандарту и обеспеченных специализированной поисковой системой.

Коэффициент кластеризации (Clustering coefficient) – величина, соответствующая уровню связности узлов в сети. Эта величина показывает, сколько ближайших соседей данного узла является ближайшим соседом друг для друга, и равна отношению реального количества ребер, которые соединяют ближайших соседей данного узла, к максимально возможному.

Кэш (Cache) – промежуточный буфер с быстрым доступом, содержащий информацию, которая может быть запрошена с наибольшей вероятностью. Доступ к данным в кэше осуществляется быстрее, чем выборка исходных данных из более медленной памяти или удаленного источника, однако ее объем существенно ограничен по сравнению с хранилищем исходных данных.

Латентно-семантический анализ, ЛСА (Latent semantic analysis, LSA) – метод обработки информации на естественном языке, анализирующий взаимосвязь между коллекцией документов и терминами в них встречающимися, сопоставляющий некоторые факторы всем документам и терминам. В основе метода латентно-семантического анализа лежат принципы факторного анализа. При кластеризации документов этот метод используется для извлечения контекстно-зависимых значений лексических единиц при помощи статистической обработки больших корпусов текстов.

Латентные факторы (Latent factors) – факторы, действие которых на изучаемый объект скрыто от исследователя. Латентные факторы вносят существенную неопределенность в ходе проведения исследований.

Лексическая единица (Lexical unit) – слово, словосочетание или лексически значимая компонента сложного слова естественного языка.

Логический поиск (Boolean search) – поиск информации с использованием логических операторов.

Масштабируемость (Scalability) – свойство системы или отдельных ее частей, характеризующее возможность системы приспосабливаться к уменьшению или увеличению ее отдельных параметров.

Межъязыковый индекс (Interlingual index) – в WordNet – индекс, всязывающий списки синонимических цепочек (синсетов), представленных на различных языках.

Метаданные (Metadata) – это данные о данных. В состав метаданных могут входить: каталоги, справочники, реестры. Метаданные содержат сведения о составе данных, содержании, статусе, происхождении, местонахождении, качестве, форматах и формах представления, условиях доступа, приобретения и использования, авторских, имущественных и смежных с ними правах на данные и др.

Метаинформация (Metainformation) – информация о способах и методах переработки информации или о том, где найти информацию.

Метапоисковая система (Metasearch system) - поисковая система, которая формирует поисковую выдачу за счет смешивания и переранжирования результатов поиска других поисковых систем.

Метод Байеса (Bayes's method) – аналитический метод, который эффективно используется при сравнении гипотез. В этом методе вероятности всех возможных исходов эксперимента объединяются с гипотезами, известными до проведения эксперимента, и затем вычисляется вероятность того, что данные гипотезы подтвердятся в ходе эксперимента.

Метод DFA (Detrended Fluctuation Analysis, DFA) – вариант дисперсионного анализа, который позволяет исследовать эффекты длительных корреляций в нестационарных рядах. При этом анализируется среднеквадратическая ошибка линейной аппроксимации в зависимости от размера отрезка аппроксимации.

Метод опорных векторов (SVM, Support vector machine) — набор алгоритмов обучения с учителем, использующихся для задач классификации и регрессионного анализа. Принадлежит к семейству линейных классификаторов. Особым свойством метода опорных векторов является непрерывное уменьшение эмпирической ошибки классификации и увеличение зазора, поэтому метод также известен как метод классификатора с максимальным зазором. Основная идея метода – перевод исходных векторов в пространство более высокой размерности и поиск разделяющей гиперплоскости с максимальным зазором в этом пространстве.

Методика аналитического исследования (Technique of analytical studies) — строго регламентированная система правил, регулирующих порядок, т.е. очередность и последовательность совершения соответствующих операций, применения методов и технологических приемов сбора, обработки и анализа информации.

Многоагентная система (Multi-agent system, MAC) – система, образованная несколькими взаимодействующими интеллектуальными агентами. Многоагентные системы могут быть использованы для решения таких проблем, которые сложно или невозможно решить с помощью одного агента. Примерами таких задач являются онлайн-торговля, ликвидация чрезвычайных ситуаций и моделирование социальных структур.

Многоагентное моделирование (Multiagent modelling, agent-based modeling) – моделирование на основе применения технологии интеллектуальных

агентов. Компьютерные модели, в которых атомарными элементами являются агенты.

Множество обучающее (Training set) – структурированный набор данных, применяемый для обучения аналитических моделей. Каждая запись обучающего множества представляет собой обучающий пример, содержащий заданное входное воздействие и соответствующий ему правильный выходной (целевой) результат.

Многоязычный информационно-поисковый тезаурус (Multilingual thesaurus) – информационно-поисковый тезаурус, содержащий лексические единицы, взятые из нескольких естественных языков и представляющий эквивалентные по смыслу понятия на каждом из этих языков. Предназначен для обработки документов (запросов) и информационного поиска с целью обмена информацией на различных естественных языках.

Модель поиска (Search Model) – модель процесса поиска информации. В основу традиционных моделей положены три главных подхода. Первый подход базируется на теории множеств. В качестве разновидностей данного подхода можно выделить следующие виды: булевская, расширенная булевская модель и нечеткие множества. Второй подход основывается на векторной алгебре. Этот подход можно представить в виде векторной, обобщенной векторной, латентно-семантической и нейросетевой моделях. Третий подход происходит из теории вероятностей – вероятностная модель.

Мониторинг (Monitoring) – специально организованное, систематическое наблюдение за состоянием объектов, явлений, процессов с целью их оценки, контроля, прогноза.

Мониторинг источников информации (Information sources monitoring) – процедура, отражающая процесс поступления информационных материалов. Результаты мониторинга могут

отображаться в табличном виде или в виде графиков, гистограмм, позволяя получать интегрированную картину динамики поступления информационных материалов по определенным тематикам.

Морфологический анализ (Morphological analysis) – экспертный метод систематизированного обзора всех возможных комбинаций развития отдельных элементов исследуемой системы. В лингвистике – определение морфологических характеристик слова.

Морфологический словарь (Morphological dictionary) – словарь, описывающий словоизменения наиболее употребимых слов выбранного языка. Как правило, в морфологическом словаре возможен поиск по части слова, что предоставляет возможность различных исследований и анализа языка.

Мотивация (Motivation) – внешнее или внутреннее побуждение субъекта к деятельности во имя достижения каких-либо целей, наличие интереса к такой деятельности и способы его инициирования, побуждения. Мотивирование составляет основу управления человеком.

Мультифрактал (Multifractal) – множество, содержащее в себе одновременно некоторое количество (зачастую бесконечное) фрактальных множеств, характеризуется спектром фрактальных размерностей.

Мультифрактальный спектр (Multifractal spectrum) – функция, которая применяется для характеристики мультифрактального множества $f(a)$ (спектр сингулярностей мультифрактала). Размер $f(a)$ равен хаусдорфовой размерности однородной фрактального подмножества L_α исходного множества L, дает доминирующий вклад в некоторую статистическую сумму.

Нейронная сеть (Neural network) – сеть, образованная взаимодействующими друг с другом нервными клетками или моделирующими их поведение компонентами.

Обзор (Review) – результат аналитико-синтетической переработки совокупности документов по определенному вопросу (проблеме, направлению), содержащий систематизированные, обобщенные и критически оцененные сведения. Может представлять собой отдельный документ, но может быть и частью другого документа: диссертации, монографии, статьи, курсовой или дипломной работы, отчета о научно-исследовательской работе и др.

Обогащение данных (Data enrichment) – процесс насыщения данных новой информацией, которая позволяет сделать их более ценными и значимыми с точки зрения решения той или иной аналитической задачи.

Обработка данных (Data processing) – процесс выполнения последовательности операций над данными.

Обработка документов (Document processing) – процесс создания и преобразования документов. Основными операциями обработки документов являются: классификация, сортировка, преобразование, размещение в базе данных и поиск.

Обучающая выборка (Training sample) – выборка, для каждого элемента которой указано, к какому классу он относится.

Омонимы (Homonyms) – разные по значению, но одинаковые по написанию и звучанию единицы языка (слова, морфемы и др.). Термин введен Аристотелем.

Омофоны [(Homophones) – слова, которые звучат одинаково, но пишутся по-разному и имеют разное значение.

Онтология (Ontology) – в информатике – попытка всеобъемлющей и детальной формализации некоторой области знаний с помощью концептуальной схемы. Используются в процессе программирования как форма представления знаний о реальном мире или его части. Основные сферы применения – моделирование бизнес-

процессов, Семантический веб, искусственный интеллект.

Опорные слова (Basic words) – главные слова, на которых строится текст.

Парадигма (Paradigm) – любая исходная концептуальная схема, модель постановки проблем и их решения. В языкознании означает систему форм одного слова, которая отражает видоизменение слова по присущим ему грамматическим категориям. Также парадигмой может называться образец типа склонения или спряжения.

Персистентность (Persistence) – неизменность, сохраняемость, устойчивость – свойство системы следовать тренду своего развития.

Пертинентность (Pertinence) – степень соответствия содержания документов информационной потребности пользователя, выражающаяся соотношением объема полезной информации к общему объему полученной информации.

Пиринговая сеть (Peer-to-peer, P2P) – компьютерная сеть, основанная на равноправии участников. В такой сети отсутствуют выделенные серверы, а каждый узел (пир – peer) является как клиентом, так и сервером. В отличие от архитектуры клиент-сервера, такая организация позволяет сохранять работоспособность сети при любом количестве и любом сочетании доступных узлов. Участниками сети являются пиры.

Поисковая машина (Search engine) – основная компонента информационно-поисковой системы. Программный модуль, осуществляющий поиск в базе данных по запросу (поисковому предписанию), заданному пользователем.

Поисковая оптимизация (Search engine optimization, SEO) – комплекс мер для поднятия позиций сайта в результатах выдачи поисковых систем по определенным запросам пользователей.

Поисковая система (search engine) – программно-аппаратный комплекс, предоставляющий возможность поиска информации. Программной частью поисковой системы является поисковая машина.

Поисковый алгоритм (Search algorithm) – математическая модель, позволяющая поисковой системе составить поисковую выдачу, релевантную запросу пользователя.

Поисковый запрос (Search query) – слово, фраза или ряд символов, по которой пользователь желает получить определенную информацию.

Полнота информации (Completeness of the information) – характеристика, определяющая количество информации, необходимое и достаточное для принятия правильного решения.

Полнотекстовая база данных (Full text database) – база данных, содержащая полные тексты документов или их частей.

Полнотекстовая поисковая система (Full-text search engine) – информационно-поисковая система, которая проводит индексирование всех слов в тексте документа (иногда за исключением стоп-слов) и учитывает порядок их расположения по отношению друг к другу.

Полнотекстовый поиск (Full text searching) - поиск текстовых документов в базе данных на основании их содержимого; автоматизированный документальный поиск, при котором в качестве поискового образа документа используется его полный текст или существенные части текста.

Пользователь информации (Information user) – субъект, обращающийся к информационной системе или посреднику за получением необходимой ему информации.

Посредничество (Betweenness) – параметр, показывающий, сколько наикратчайших путей

проходит через узел. Эта характеристика отражает роль данного узла в установлении связей в сети.

Предметная рубрика (Subject heading) – наименование классификационного признака однородных объектов библиографической или информационной деятельности.

Профайл (Profile) – совокупность величин определяющих параметров некоторого объекта или технологического процесса, описывающих и характеризующих этот объект или технологический процесс.

Ранжирование (Ranking) – упорядочение результатов поиска – отклика поисковой системы по некоторым критериям, например, по дате публикации документов или по релевантности.

Редукционизм (Reductionism) – методологический принцип, основывающийся на возможности объяснения сложного на основе законов простого. Редукционизм абсолютизирует принцип редукции – сведния сложного к более простому и высшего к низшему.

Редукция (Reduction) – упрощение, сведение сложного произвольного процесса к более простому, более доступному для анализа и решения; логико-методологический прием, заключающийся в сведении в процессе исследования одного явления к другому, одной задачи (или проблемы) к другой с целью упрощения.

Релевантность (Relevance) – степень соответствия запроса и найденного, то есть уместность результата. В более общем смысле не только оценка степени соответствия, но и степени практической применимости результата, а также степени социальной применимости варианта решения задачи.

Репозиторий, хранилище (Software repository) — место, где хранятся и поддерживаются данные чаще всего данные в виде файлов.

Репрезентативная выборка (Representative sample) – истинное отражение родительской популяции, т.е. выборка, которая имеет такое же распределение относительных характеристик, что и генеральная совокупность; выборка, имеющая такое же распределение относительных характеристик, что и генеральная совокупность.

Репрезентативность (Representativeness) – свойство выборки отражать характеристики изучаемой генеральной совокупности.

Репрезентативность информации (Representativity of information) – представительность информации, достаточная для обоснования решения, ради которого она собрана. Если выборка репрезентативна, то по ее свойствам можно судить о генеральной совокупности; если выборка произведена неправильно, говорят об ошибке репрезентативности.

Реферирование (Referencing) – процесс извлечения основного содержания сообщения или документа с использованием совокупности интеллектуальных процедур, основанных на различных методах. Может осуществляться как экспертом, так и с применением средств автоматизации.

Рубрикатор (Rubricator) – классификационная таблица иерархической классификации, содержащая полный перечень включенных в систему классов и предназначенная для систематизации информационных фондов, массивов и изданий, а также для поиска в них.

Рубрикация (Rubrication) – процесс распределения документов по по разделам (рубрикам).

Саморазвивающаяся система (Self-developing system) – кибернетическая адаптируюшаяся (динамическая) система, которая самостоятельно вырабатывает цели своего развития и критерии их достижения, изменяет свои параметры, структуру и другие характеристики в заданном направлении.

Самоорганизация (Self-organization) – процесс упорядочения элементов одного уровня в системе за счет внутренних факторов, без внешнего специфического воздействия.

Самоподобный объект (Self-similar object) – объект, в точности или приближенно совпадающий с частью себя самого (то есть целое имеет ту же форму, что и одна или более частей). Инвариантность относительно изменения шкалы является одной из форм самоподобия, при которой при любом приближении найдется по крайней мере одна часть основной фигуры, подобная целой фигуре.

Семантика (Semantics) – раздел языкознания и логики, исследующий проблемы, связанные со смыслом, значением и интерпретацией лексических единиц. В программировании – система правил истолкования отдельных языковых конструкций. Семантика определяет смысловое значение предложений алгоритмического языка

Семантическая информация (Semantic information) – информация, содержащаяся в высказывании и передаваемая (переводимая) через значения единиц речи (языка).

Семантическая сеть (Semantic network) –структура данных, состоящая из узлов, соответствующих понятиям, и связей, указывающих на взаимосвязи между узлами, отношения между ними.

Семантические модели данных (Semantic data model) – представляют собой средство представления структуры предметной области. Используют общий набор понятий и отличаются конструкциями, применяемыми для их выражения, полнотой отражения понятий в модели, удобством использования при разработке информационных систем.

Семантический информационный барьер (Semantic information barrier) – информационный барьер, обусловленный несовпадением толкований

одних и тех же слов, терминов и символов разными людьми.

Семантический поиск (Semantic search) – процесс разыскания информационных сообщений по их смыслу, содержанию.

Сетевая мобилизация (Network mobilization) – процесс объединения усилий участников виртуального пространства социальных сетей для решения некоторых проблем, возникающих (или имеющих место) в реальном мире.

Синергетика (Synergetic) – междисциплинарное направление научных исследований, задачей которого является изучение природных явлений и процессов на основе принципов самоорганизации систем (состоящих из подсистем).

Синергетический подход (Synergetic approach) – совокупность принципов, основой которой является рассмотрение объектов как самоорганизующихся систем.

Синонимы (Synonyms) – слова близкие или тождественные по своему значению, выражающие одно и то же понятие, но различающиеся или оттенками значения, или стилистической окраской, или и тем и другим.

Система клеточных автоматов (Cellular automata system) – совокупность математических объектов, представляющих собой однородную сетку, каждая клетка которой (клеточный автомат) может находиться в одном из возможных состояний. Состояния клеток синхронно обновляются на каждом шагу моделирования соответствии с установленными правилами перехода, в общем случае таких правил может быть бесчисленное количество, соответствующее количеству подмножеств счетного множества.

Скейлинг (Scaling) – масштабная инвариантность, самоподобие. Это свойство применяется, в частности,

для представления функции двух переменных как функции одной.

Скрытый веб (Deep Web) – множество веб-документов, которые не охватываются традиционными для обычного веб-пространства информационно-поисковыми системами. Как правило, эти веб-документы доступны в сети Интернет, однако выйти на них невозможно, если не знать точного адреса. К этим ресурсам относятся и динамично формируемые веб-документы, содержание которых хранится в базах данных и доступно только по запросам пользователей.

Словоформа (Wordform) – форма слова, полученная из именительного падежа существительного, или инфинитива глагола с помощью склонения или спряжения.

Словарь стоп-слов (Dictionary of stop words) – содержит специфичные слова, которые используются часто и поэтому бесполезны в качестве условий поиска.

Словарь частотный (Frequency word-book) – словарь, содержащий перечень слов данного языка, расположенных по степени их употребительности.

Словоформа (Word form) – грамматическая форма слова; термин, обозначающий конкретное слово в конкретной грамматической форме.

Сниппет (Snippet) – часть текста, отрывки веб-страницы, которая содержит слова поискового запроса, выводящаяся информационно-поисковой системой в результатах поиска по данному запросу.

Событие (Event) – изменение свойств, параметров исследуемого объекта или процесса в определенный момент времени, зарегистрированное пользователем.

Сообщение (Message) – совокупность данных, содержащих какие-либо сведения, предназначенные для передачи по каналу связи от источника к потребителю (получателю) сообщения. Сообщением может быть число, знак, текст, изображение и т.д. При передаче С.

кодируются, т.е. преобразуются в сигналы на входе канала, а затем декодируются, принимая форму, доступную для восприятия.

Социальная информация (Social information) – совокупность знаний, сведений, данных и сообщений, которые формируются и воспроизводятся в обществе и используются индивидами, группами, организациями, различными социальными институтами для регулирования социального взаимодействия, общественных отношений и процессов.

Социальная сеть (Social network) – структура, состоящая из узлов (которыми являются социальные объекты) и связей между ними. Объектами сетей могут быть предприятия, люди, научные учреждения, интернет–ресурсы и т.д.

Старение информации (Ageing of information) – свойство информации утрачивать со временем свою практическую ценность, обусловленное изменением состояния отображаемой ею предметной области.

Стемминг (Stemming) – процесс нахождения основы слова для заданного исходного слова. Основа слова необязательно совпадает с морфологическим корнем слова.

Стоп-слова (Stop words) – слова, исключаемые из индекса системы или запроса пользователя. К стоп-словам обычно относятся предлоги, междометия и другие сочетания, которые не несут содержательного смысла.

Субъект (Subject) – активный компонент системы, обычно представленный в виде пользователя, процесса или устройства, который может явится причиной потока информации от объекта к объекту или изменения состояния системы.

Сценарий (Script, scenario) – план выполнения процесса; определяет последовательность команд, которая указывает программе, как и в каком порядке

выполнять ту либо иную процедуру. Составляется на языке сценариев.

Сценарный анализ (Scenario analysis) – методика анализа данных, где используется набор детальных описаний последовательности событий, которые с прогнозируемой вероятностью могут привести к желаемому или планируемому конечному результату или к возможным исходам, при рассматриваемых сценаристом различных вариантах развития исследуемого процесса. Основными приложениями сценарного анализа являются стратегическое планирование и управление, а также оценка рисков и прогнозирование.

Тезаурус (Thesaurus) – словарь, описывающий лексическую семантику, в котором слова сгруппированы в соответствии с понятийной классификацией с заданными смысловыми отношениями.

Текстовая база данных (Text database) – база данных, записи в которой содержат (главным образом) текст на естественном языке.

Текстовые данные (Text data) – последовательность символов, соответствующих в том или ином наборе символов буквам алфавита и знакам препинания.

Текстовый интерфейс (Character based interface) – интерфейс пользователя, в котором вся информация на экране представлена в виде текста.

Текстовый корпус (Text corpus) – массив текстов, собранных в соответствии с определенными принципами, размеченных по определенному стандарту и обеспеченных специализированной поисковой системой. В некоторых случаях текстовым корпусом первого порядка называют произвольное собрание текстов, объединенных каким-то общим признаком. Разработкой, созданием и использованием текстовых

(лингвистических) корпусов занимается специальный раздел языкознания – корпусная лингвистика.

Текстовый файл (Text file) – файл, содержащий текстовые данные, как правило, организованные в виде строк. Текстовый файл, как и прочие файлы, хранится в файловой системе. Текстовый файл может содержать как форматированный, так и неформатированный текст.

Терм (Term) – слово или устойчивое словосочетание. Понятие «терм» как «символьное выражение» широко используется в математической логике.

Термин (Term) – слово либо словосочетание, обозначающее понятие, применяемое в науке, технике, искусстве и т.д. Совокупность терминов образует терминологию.

Терминологическая система (Terminological System) – знаковая модель определенной области знаний или деятельности. Элементами терминосистемы служат лексические единицы (слова и словосочетания) определенного языка.

Тональность (Tonality) – эмоциональное отношение автора высказывания к некоторому объекту, выраженное в тексте. Эмоциональная составляющая, выраженная на уровне лексемы или коммуникативного фрагмента, называется лексической тональностью (или лексическим сентиментом). Тональность всего текста в целом можно определить как функцию (в простейшем случае сумму) лексических тональностей составляющих его единиц (предложений) и правил их сочетания.

Тренд (Trend) – аналитическое или графическое представление изменения переменной во времени, полученное в результате выделения регулярной составляющей динамического ряда, характеризует существующую динамику развития процесса в целом. Случайная составляющая отражает случайные колебания или шумы процесса.

Фазовое пространство (Phase space) – множество всех состояний системы в фиксированный момент времени. Каждому возможному состоянию системы соответствует точка фазового пространства. Сущность понятия фазового пространства заключается в том, что состояние сколь угодно сложной системы представляется в нем одной единственной точкой, а эволюция этой системы – перемещением этой точки.

Фактор (Factor) – источник воздействия, приводящего к изменению значений переменных модели некоторой системы; движущая сила какого–либо процесса или явления. Часто термины «фактор» и «переменная» (признак, показатель) отождествляются, что не всегда справедливо.

Флективные языки (Inflected languages) – языки, в которых в выражении грамматических значений ведущую роль играет флексия (окончание).

Формализация (Formalization) – описание теорий, осмысленных предложений и т. п. формальными средствами, прежде всего символами математики и математической логики.

Фразеология (Phraseology) – раздел языкознания, изучающий лексико-семантическую сочетаемость слов языка.

Фрактал (Fractal) – бесконечно самоподобный (точно или приближенно) объект (множество), каждая часть которого повторяется при уменьшении масштаба. Размерность Хаусдорфа-Безиковича такого объекта должна быть нецелой, поэтому фрактал самоподобен, обратное не обязательно. Возможно и такое определение: фрактал – самоподобное множество нецелой размерности.

Фрактальный анализ (Fractal analysis) – метод моделирования данных с помощью теории фракталов, заключающийся в исследовании фрактальной размерности и других фрактальных свойств сигналов, наборов данных, объектов.

Функционал (Functional) – переменная величина, заданная на множестве функций, зависящая от одной или нескольких функций. Функция, аргументы которой также представляют собой функции некоторых переменных.

Функция (Function) – 1) действие, осуществляемое в рамках процесса; 2) содержание действий, выполнение которых возлагается на элемент системы при заданных требованиях, условиях и ограничениях; 3) в математике – правило, по которому каждому значению одной или нескольких переменных, называемых аргументами, ставится в соответствие только одно значение переменной, называемой функцией.

Хранилище информации (Repository of information) – совокупность информационных систем, включая базы данных и справочники, которые реализуют функциональность по описанию метаданных, сбору, очистке, обогащению, консолидации первичной информации с транзакционных систем, а также по визуализации (построению витрин) данных.

Целостность (Integrity) – относительная независимость системы от среды и от других аналогичных систем. Целостность выражает интегрированность, самодостаточность, автономность этих объектов, их противопоставленность окружению, связанную с их внутренней активностью; характеризует их качественное своеобразие, обусловленное присущими им специфическими закономерностями функционирования и развития.

Ценность информации (Information value) – свойство информации, определяемое ее пригодностью к практическому использованию в различных областях целенаправленной деятельности человека.

Цепь Маркова (Markov chain) – марковский процесс с дискретным временем и конечным или счетным множеством состояний.

Эмерджентность (Emergence) – наличие у какой-либо системы особых свойств, не присущих ее подсистемам и блокам, а также сумме элементов, не связанных особыми системообразующими связями; несводимость свойств системы к сумме свойств ее компонент; синоним – «системный эффект».