АКАДЕМИЯ ПРАВОВЫХ НАУК УКРАИНЫ

Научно-исследовательский центр правовой информатики

Д.В. ЛАНДЭ, В.Н. ФУРАШЕВ,
С.М. БРАЙЧЕВСКИЙ, А.Н. ГРИГОРЬЕВ

ОСНОВЫ МОДЕЛИРОВАНИЯ И ОЦЕНКИ ЭЛЕКТРОННЫХ ИНФОРМАЦИОННЫХ ПОТОКОВ

Киев 2006
ООО "Инжиниринг"

УДК 681.3
ББК 32.973.26-018.2.75
Л 22

Рекомендовано в печать
Ученым советом Научно-исследовательского
центра правовой информатики Академии правовых наук Украины
(протокол № 7 от 01.08.2006)

Рецензенты

Н.Я. ШВЕЦ - доктор экономических наук, Заслуженный деятель науки и техники Украины, член-корреспондент АПрН Украины, профессор
А.Г. ГРЕБЕННИКОВ – доктор технических наук, Заслуженный работник образования Украины, профессор
С.Н. ДАНИЛЯК – доктор технических наук

Л 22 **Ландэ Д.В., Фурашев В.Н., Брайчевский С.М., Григорьев А.Н.**
Основы моделирования и оценки электронных информационных потоков: Монография. – К.: Инжиниринг, 2006. – 176 с.
ISBN 966-95147-6-2.

В работе рассматриваются теоретические, методологические и технологические вопросы моделирования и оценки электронных информационных потоков. В качестве базы для апробации предложенных подходов рассматриваются потоки новостей в сети Интернет. Приведены математические модели информационных потоков, рассматриваются фрактальные свойства информационного пространства, концепции поиска информации. В монографии рассмотрены основные компоненты концепций Семантического Web, Web второго поколения, ориентированные на обеспечение обобщенного доступа к сетевому контенту. Большое внимание уделено описанию таких практически важных вопросов, как принципы построения систем контент-мониторинга, определение тональности сообщений, выявление дубликатов.

Рассчитана на широкий круг читателей.

УДК 681.3
ББК 32.973.26-018.2.75

ISBN 966-95147-6-2
ООО "Инжиниринг"
Заказ №
Тираж 500 экз.

Содержание

Введение

Принятие решений в народном хозяйстве, экономике, политике, научно-технической и социальной сферах, как известно, базируется на процессах сбора, анализа и синтеза информации, то есть всегда нуждается в серьезной информационной поддержке. Удовлетворение информационных потребностей в настоящее время является обязательной предпосылкой осуществления инновационных преобразований. Вместе с тем сложность получения информации влияет на оперативность и качество принятия решений. Поэтому задачу обобщения, интеграции современных информационных потоков можно считать наиболее актуальной в условиях стремительного развития экономических, политических и общественных процессов.

Современным информационным потокам присущи многовариантность и многофакторность. Среди решающих факторов можно определить время, уменьшение влияния которого, то есть задержки в принятии решений, позволяет, в частности, экономить производственные ресурсы за счет принятия обоснованных решений, получать экономический эффект.

Надо отметить, что наряду с ростом объемов информации возрастает и количество информационных источников. Одним из классов таких источников выступает информационная составляющая сети Интернет. В рамках данной работы информационные потоки в Интернет рассматриваются как полигон, информационный корпус, динамика и объемы которого, в частности, обусловили на это время появление проблемы ориентации в его новостной части.

Поэтому как база для решения актуальной задачи интеграции современных информационных потоков выбрана именно новостная составляющая сети Интернет, динамика и объемы которой на сегодня достигают больших значений. Именно бурное развитие Интернет в последнее время породило ряд специфических проблем, связанных, в первую очередь, с быстрым ростом объемов данных, подлежащих хранению и обработке.

В начале существования World-Wide Web на небольшом количестве Web-сайтов публиковалась информация отдельных авторов для относительно большого количества посетителей. Сегодня ситуация резко изменилась. Сами посетители Web-сайтов активно участвуют в создании контента, что привело к резкому росту объема и динамики информационного пространства.

Сегодня в Интернет уже существует доступная для экспериментов 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, т.е. за один год количество открытой, доступной простому пользователю информации из Интернет удвоилось. По данным службы Web Server Survey, в августе 2006 года количество Web-сайтов превысило 94 миллиона (рис. 1). Таким образом, приведенные данные подтверждают экспоненциальный характер роста объемов данных в Сети.

Этот рост сопровождается рядом проблем [18], [23], таких как:

- непропорциональный рост уровня информационного шума;
- засилье паразитной информации (невостребованной, получаемой в качестве несанкционированных "приложений");
- слабая структурированность информации;
- многократное дублирование информации.

Web-пространству к тому же присущи такие недостатки, как обилие «информационного мусора», невозможность гарантирования целостности документов, практическое отсутствие возможности смыслового поиска, ограниченность доступа к «скрытым» ресурсам (Deep Web).

Над решением названных проблемам работают многочисленные коллективы ученых и специалистов во всем мире, в частности, консорциум W3C, где реализуется концепция Семантического Web [19], [42]. Наряду с этой концепцией, революционный прорыв обещает дать более общий подход, а именно Web-2.0 (http://www.web2con.com/), или как его называют, “Web второго

поколения", который предполагает реализацию концепции семантического Web, включая многоуровневую поддержку метаданных, новые подходы к дизайну и соответствующему инструментарию, технологию глубинного анализа текстов (Text Mining), а также идеологию Web-сервисов, базируясь при этом на информационных ресурсах, накопленных в WWW первого поколения.

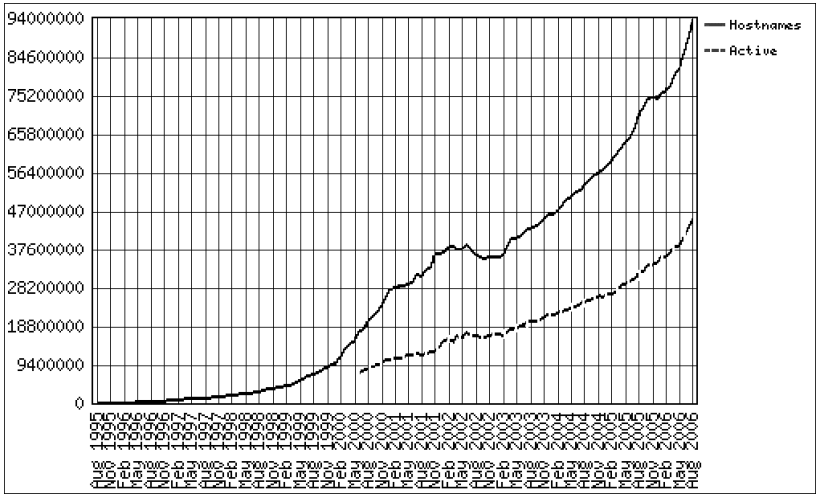


Рис.1. Динамика роста количества Web-сайтов

В настоящее время в связи с развитием информационных ресурсов сети Интернет документальное информационное пространство развилось до уровня, требующего новых подходов. Рост объемов информации и скорости ее распределения фактически породил понятие информационных потоков [1]. Вместе с тем существующий математический аппарат и инструментальные средства уже не всегда способны адекватно отражать ситуацию, речь идет не столько о конечных массивах документов, сколько о динамичных документальных информационных потоках.

Сегодня есть основания полагать, что определенного переосмысления требует само понятие информации, в частности, его взаимосвязь с понятием «знания», ставшим особенно популярным в последние годы. Часто

употреблявшийся ранее в теории искусственного интеллекта и впоследствии основательно забытый термин «преобразование информации в знания», похоже, вновь начинает вызывать интерес.

Этому в значительной мере способствовали чисто прикладные успехи в машинной обработке потоков данных, содержащих документы, не только составленные на разных языках, но и относящиеся к различным социокультурным контекстам. Ясно, что в таком случае обработка потока данных (т. е. информации в чистом виде), какой бы она ни была, не предполагает активного использования содержания документов. Теоретическое осмысление подобных ситуаций наводит на мысль о том, что собственно «знания» представляют собой некую надстройку над информационными потоками, определяемую в конечном счете наличием устойчивых связей между определенными информационными элементами (данными).

Практика показывает, что информация может вполне успешно обрабатываться вне зависимости от того, какой смысл в нее заложен. В связи с этим вновь возник интерес к подходам, основанным на понимании информации как меры упорядоченности некоторой системы и, соответственно, к статистическим методам ее обработки. Многие ученые и ведущие участники информационного рынка возвращаются к истокам теории информации, понятиям энтропии, теории Шеннона, уравнениям Больцмана и др.

При этом оказалось, что многие задачи, возникающие при работе с информационными потоками, имеют немало общего с задачами статистической физики и гидродинамики и могут решаться одними и теми же методами. Это обстоятельство открывает широкие перспективы применению мощного аппарата современной физики к решению теоретико-информационных задач.

С другой стороны, признание того, что извлечение из информационных потоков знаний в обычном смысле слова является самостоятельной проблемой, которая должна решаться с использованием новых подходов, несомненно будет способствовать развитию традиционной методологии.

Если информация может обрабатываться независимо от содержательного аспекта, то обратное не верно. В любом случае именно информация является своего рода «субстратом знаний».

Поэтому теория информации, которая ранее находила свое основное реальное применение в области техники передачи информации, сейчас становится полезной и для анализа смысловых текстовых потоков. Энтропия информационного пространства с помощью осмысленного анализа уменьшается постепенно, но чем более комплексный этот анализ, тем заметнее переход от хаоса к порядку.

Проблема «знаний», скорее всего, никогда не будет сведена к какому-либо комплексу задач, которые можно было бы окончательно решить чисто технологическим путем. Напротив, она, видимо, потребует серьезных исследований в различных направлениях, в том числе и на достаточно высоком теоретическом уровне.

Одним из центральных вопросов в этом плане, на наш взгляд, является отношение информационного и семантического пространства, чему, как правило, уделяется неоправданно мало внимания. В литературе их часто даже отождествляют, без всякого на то основания. То, что эти две категории никоим образом не тождественны, с очевидностью вытекает из различия их природы: информационное пространство образуют данные, физически записанные на тех или иных носителях, тогда как семантическое пространство порождают комплексы абстрактных понятий, связанных с субъективными оценками, даваемыми человеком. Наиболее естественным представляется определить сетевое семантическое пространство как множество единиц смысла, актуальных в данном социокультурном контексте и представленных в Сети. Под единицей смысла мы, как обычно, понимаем элементарную категорию, позволяющую нам строить субъективные оценочные суждения о вещах и процессах, относящихся к окружающему нас миру.

В реальной жизни между ними, безусловно, существует вполне определенная связь, но нахождение этой связи, по-видимому, представляет собой весьма нетривиальную задачу.

Глубину и важность проблемы понимания соотношения информационного и семантического пространства проиллюстрируем на примере автоматического реферирования текстового массива, содержащего документы, составленные на разных языках. Возможен ли алгоритм, позволяющий выделить из произвольного документа информационно значимые фрагменты, «не зная» языка, на котором этот документ создан? Разумеется, речь не идет об идентификации языка документа с последующим подключением соответствующих словарей, наборов грамматик и т. п.

Оказывается, такой алгоритм возможен, если только входной поток удовлетворяет законам Зипфа, т.е. создан человеком. Более того, он успешно реализован авторами в системе InfoStream® [27]. И это порождает философский вопрос: в какой мере понятие «информация» связана с понятием «смысл», и связаны ли эти понятия вообще? По крайней мере, в общепринятом понимании.

Действительно, практика показывает, что определенное количество информации может быть передано и надлежащим образом обработано (с чем конечный потребитель, прочитав реферат и сравнив его с оригиналом, согласится), но при этом в процессе обработки смысл, возможно содержащийся в ней, никак не учитывался. Более того, мы можем даже не знать, имеют ли вообще какое-либо значение в семантическом отношении последовательности символов, составляющих документ («Глокая куздра штеко будланула бокра…»).

Самое интересное при этом заключается в интерпретации полученных результатов. Без привлечения методов искусственного интеллекта, объемных средств семантической формализации, даже экспертов как таковых, с использованием только частотных методов могут быть получены содержательные, семантически наполненные результаты. Возникает ощущение, что для полноценной работы с информацией вполне достаточно структурно-лингвистического уровня.

При желании можно было бы возразить, что поскольку смысл сопряжен с понятиями, то есть знаками вещей, а не со знаками знаков – словами, он, вообще говоря, от языка не зависит, и этот факт позволяет осуществлять переводы текстов. Парадокс, однако, в том, что алгоритм, о котором идет речь, переводом (как бы мы его ни понимали) как раз и не занимается: он просто «не обращает внимания» на содержание документа.

В качестве других примеров можно привести автоматическое выявление взаимной связи понятий, автоматическую группировку (кластеризацию) связей для выявления наиболее важных из них, автоматическое выявление их "окраски", в простейшем случае - определение принадлежностей взаимосвязей к положительным (группирующим) или отрицательным (антагонистическим).

Вместе с тем потребитель все же хочет в конечном счете получить нечто осмысленное. Поэтому полное игнорирование семантических аспектов в информационных технологиях было бы ошибочным. Видимо, оптимальный путь состоит в том, чтобы более адекватно оценить функциональную роль и значение семантического уровня информационных процессов. Поэтому и возникает вопрос о природе связи информационного и семантического пространств.

Очевидно, что для соотнесения элементов информационного и семантического пространств необходим некий промежуточный модельный уровень обработки текстовых данных. При этом должны быть определены «правила чтения», с помощью которых формальная система (набор структурных элементов текста) преобразуется в систему содержательную (осмысленное сообщение). Более того, эти правила должны быть встроены в некую компьютерную программу. И здесь возникает серьезная сложность. В реальной жизни такие правила никогда не формализуются. Человек постигает их годами, активно действуя в определенном социокультурном контексте, постоянно общаясь с другими людьми. Причем различные контексты порождают различные «правила чтения», которые, к тому же, изменчивы во времени.

Очевидно, в наше время не существует единого способа научить таким правилам машину, а без этого, в свою очередь, невозможно добиться того, чтобы

она, обрабатывая текст, учитывала его содержательный аспект предсказуемым образом.

Таким образом, с информационным пространством оказывается сопряжено не только семантическое пространство, которое может быть доступно нашему интеллекту, но и пространство формальных «правил чтения», позволяющих производить заданный набор операций.

“Обратной стороной медали” является тот несомненный факт, что информационное пространство, в конечном счете, порождается семантическим. Действительно, возникновение информационных потоков можно представить себе как генерацию и движение наборов данных, ассоциированных с определенным сообщением, понимаемым как некоторый смысловой блок. Конечно, одному сообщению может соответствовать произвольное число отдельных наборов данных (о важном международном событии напишут все медийные средства). Таким образом, характеристики информационного пространства изначально определяются структурой сетевого семантического пространства, причем здесь мы имеем право говорить о структуре, поскольку сообщения отражают события реального мира, который, вероятно, все же в какой-то мере упорядочен.

Эпиграфом к тематике поиска информации могли бы послужить слова персонажа кинофильма «Уолл-стрит»: скажи мне то, чего я не знаю! Попытки технологического развития в рамках современной теории информационного поиска сегодня очень часто не улучшают, а ухудшают ситуацию. Например, совершенствование технических аспектов информационно-поисковых систем лишь приводит к увеличению объемов релевантных наборов, которые зачастую не пригодны к употреблению.

Современные технологии позволяют осуществлять невероятно изощренные операции над данными, но чем эффективнее они применяются, тем менее приемлемым оказывается результат.

Очевидно, следует признать, что изначальная парадигма поисковых систем, сформированная десятки лет тому назад, уже не отвечает реальной ситуации.

Таким образом, возникает задача поиска новых принципов, в рамках которых оказалось бы возможным проектирование качественно новых систем обработки больших и динамичных объемов данных.

Поисковые машины следующих поколений должны будут лучше классифицировать информацию и нагляднее представлять ее. В будущем поиск не должен ограничиваться лишь обработкой введенных ключевых слов. Имеет смысл реализация перехода к концепции навигации в информационных потоках, как к распределенному во времени интерактивному процессу локализации отдельных семантических секторов в общем информационном пространстве.

Системы должны будут отслеживать интересы пользователей, делая поиск более целенаправленным. Можно предположить, что новые поисковые машины будут находить опубликованные в сети текстовые, аудио- и видеоматериалы, которые в настоящее время недоступны.

Надежды, возлагавшиеся в свое время на идею последовательного уточнения поиска («искать в найденном», «показать подобное» и т. п.) не оправдались по двум причинам. Во-первых, интересующий потребителя документ может просто не оказаться в первичной выборке, в силу чего последующие итерации теряют смысл, а во-вторых, составление уточняющего запроса, качественно отличающегося от исходного, представляет собой отнюдь не простую задачу, прямо скажем, непосильную для рядового пользователя.

Мы не ищем в сетях то, что и так знаем – в этом нет никакого смысла. Нам нужно что-то, чего мы не знаем, и мы лишь пытаемся объяснить машине, в каком сегменте информационного пространства это «что-то», по нашему мнению, должно находиться. Указание набора слов, с помощью которых этот сегмент можно локализовать, оказывается отнюдь не самым совершенным способом достижения поставленной цели.

Таким образом, резюмируя приведенные выше рассуждения, выскажем предположение о том, что современные информационные технологии готовы к пересмотру принципов обеспечения доступа к сетевым данным, который условно можно назвать переходом от информационного поиска к сетевой навигации.

1. Новостной Web

Эффективный анализ новостных информационных потоков в Интернет, построение систем синдикации новостей невозможны без некоторых сведений о структуре новостного Web-пространства [18]. Это пространство формируется динамичными потоками сообщений, публикуемых на Web-сайтах средств массовой информации, информационных агентств, отдельных организаций. Чем, например, может быть полезно знание о структуре новостного Web-пространства на практике? Во-первых, это даст возможность выявления первоисточников информации [21], [26], 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, «мусорных» источников. Кроме того, для оперативного нахождения актуальной информации корректная модель может способствовать нахождению действительно полезных первоисточников и служб интеграции информации.

Если для обычного Web-пространства уже признана модель «галстука-бабочки», представленная в работах А. Бредера и его коллег [43], [44], [45],то публикации об архитектуре новостного Web-пространства до настоящего времени не известны. Можно было бы попросту применить вышеназванную модель А. Бредера к новостной составляющей Web-пространства, однако такой подход нельзя считать корректным по ряду причин:

- новостные потоки характеризуются динамикой [51], 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 «скрытого» Web, т.е. тех информационных Web-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 Web-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 Web-пространства наибольшее внимание должно уделяться именно Web-сайтам, на которых публикуются новостные сообщения, а не отдельным Web-страницам или самим сообщениям.

В качестве экспериментальной базы для построения модели новостного Web-пространства авторами использовался достаточно мощный информационный корпус - ретроспективная база данных системы контент-мониторинга InfoStream [37]. Система InfoStream применяется для решения задач автоматизированного сбора новостной информации с открытых Web-сайтов и обеспечения доступа к ней в поисковых режимах. Эта разработанная в Информационном центре «ЭЛВИСТИ» система в настоящее время охватывает свыше 2500 источников, а ретроспективные базы данных системы представляют собой корпус объемом более 30 млн. документов. Для построения модели использовалась база данных новостных сообщений за февраль 2006 года объемом около 760 тыс. документов. Для каждого из источников был составлен запрос в следующем виде:

<код источника>#<шаблон для поиска>[#<шаблон для поиска>…#<шаблон для поиска>],

совокупность которых была объединена в конфигурационном файле, фрагмент которого представлен ниже:

srd00001#УКРОП#ukrop.com

srd00002#BBC#bbc.co.uk

srd00003#Champion.com.ua#champion.com.ua

srd00004#Crashes.ru#crashes.ru

srd00006#"Немецкая волна#Deutsche Welle#dwelle.de

srd00007#GazetaSNG.ru#gazetasng.ru

srd00008#idNews.com.ua#idnews.com.ua

srd00011#InoPressa#Инопресса#inopressa.ru

srd00012#Internet.ru#internet.ru

srd00013#K2Kapital#k2kapital.com

srd00014#KPNews.com#kpnews.com

srd00015#Lenta.Ru#Лента.ру#lenta.ru

srd00016#MIGnews.com#mignews.com

В результате специальной обработки такого пакета запросов для каждого сообщения, относящегося к определенному источнику - Web-сайту были выявлены исходящие ссылки на другие источники (ссылки на собственный источник исключались). Было выявлено, что исходящие контекстные ссылки присутствовали на 264942 сообщениях с 1531 Web-сайта. Общее же количество Web-сайтов, участвующих в процессе взаимных ссылок, составило 1863. Было выявлено также 54 источника, не входящих в этот список, т.е. тех, на которые не вела ни одна из контекстных ссылок и сообщения которых не ссылались ни на одни из исследуемых Web-сайтов. Такие Web-сайты («абсолютные острова») были вынесены за рамки модели.

Ниже приведен список Web-сайтов, обладающих максимальным количеством исходящих ссылок:

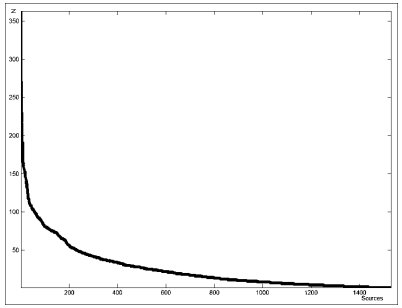
Web-сайт	**Количество ссылок**
RAMBLER	363
VLASTI.NET	271
RosInvest	270
"Обозреватель"	231
ИА "REGNUM"	217
Деловая пресса"	202
“Россия-Он-Лайн"	193
"Оглядач"	191
RNews	183

Web-сайт	Количество ссылок
Fin.org.ua	166
PRESIDENT.ORG.UA	164
"Промышленно-торговые новости"	160
"4 ВЛАДА"	159
"Украина промышленная"	156

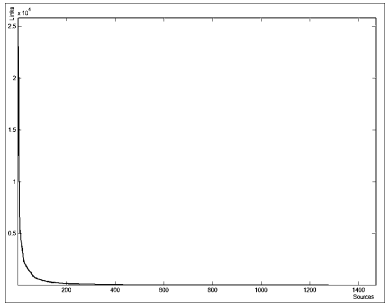
Также было получено распределение новостных Web-сайтов по количеству входящих ссылок. Всего за февраль ссылки указывали на 1470 источников (без самоцитирования). Оказалось, что на 100 источников ведет свыше 80% ссылок. На рис. 2 представлены графики ранжированных распределений новостных Web-сайтов по количеству исходящих и входящих ссылок. Следует обратить внимание на то, что второй график значительно круче первого, это говорит о большей равномерности распределения множества исходящих ссылок, чем входящих.

Ниже приведен начальный фрагмент ранжированного списка источников, на которые ведет максимальное количество ссылок:

Web-сайт	Количество ссылок
РИА "Новости"	25827
ИА "Интерфакс"	23765
ИА "REGNUM"	22354
УНИАН	17847
ИТАР-ТАСС	14157
"Reuters"	10754
"Газета.Ru"	9354
РИА "РосБизнесКонсалтинг"	7653
УНІАН	5472
"Lenta.Ru"	5223
ИА "Интерфакс-Украина"	5073
ИА "Росбалт"	4815
"proUA"	4814
НТВ	4395



Распределение по количеству исходящих ссылок



Распределение по количеству входящих ссылок

Рис. 2. Распределения источников по количеству ссылок

Кроме того, были выявлены источники, на которые не ссылаются, но которые обладают исходящими ссылками (393) и цитируемые источники, не ссылающиеся ни на кого (332).

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 новостном Web-пространстве. Это объясняется подбором источников для сканирования, в число которых не входят многие новостные интеграторы.

Выявление дублирующихся по содержанию новостных сообщений, охватываемых системой InfoStream, выполняется на основе лингвостатистических методов, заключающихся в выявлении наиболее весомых слов в документах, которые выступают своеобразными ключами. Опыт показал, что в русско- и украиноязычных потоках новостей совпадение наиболее весомых 6 ключевых слов из сообщений с более чем 95% вероятностью свидетельствует о содержательном дублировании.

Следует отметить, что применение более «мягкого» критерия к множеству отобранных ключевых слов позволяет реализовать режим «поиска подобных документов».

Было проведено исследование соотношения дублирующихся и оригинальных сообщений (см. гл. 7), которые привели к неожиданному результату. Оказалось, что количество оригинальных сообщений и их содержательных дублей, охватываемых системой InfoStream в 2005 году, почти в точности совпало.

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 сообщение многократно дублируется на сайтах и в СМИ) или при определении эффективности PR-кампаний (подсчет републикаций пресс-релизов и др.). При построении модели исследовался уровень дублирования для документов с одних Web-сайтов, имеющих ссылки на другие сайты-источники.

В результате проведенных исследований была принята модель новостного Web-пространства, которая представлена на рис. 3. Эта модель включает такие зоны:

- входной полуостров. Web-сайты, которым соответствуют менее порогового значения входящих ссылок и любое превышающее пороговое количество исходящих ссылок (таких Web-сайтов оказалось 312 или 16,7%);
- выходной полуостров. Web-сайты, которым соответствуют менее порогового значения исходящих ссылок и любое превышающее пороговое количество входящих ссылок (таких Web-сайтов оказалось 513 или 27,5%);
- остров. Web-сайты, которым соответствуют менее порогового значения исходящих и входящих ссылок (таких Web-сайтов оказалось 358 или 19,3%);
- ядро, состоящее из трех областей: входной, выходной и коммуникационной зоны (таких Web-сайтов оказалось 680 или 36,5%). Зона ядра характеризуется средними и большими значениями уровней

исходящих и входящих связей, однако, как видим, допускает ранжирование по уровню этих коммуникаций.

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 [17]. 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.

Например, для разделения области ядра на входную, выходную и коммуникационную зоны можно рассмотреть ранжированный график логарифма отношения количества исходящих и входящих ссылок для каждого из источников этой области (рис. 4). Центральная зона этого графика соответствует коммуникационной, левая – выходной, а правая – входной зоне.

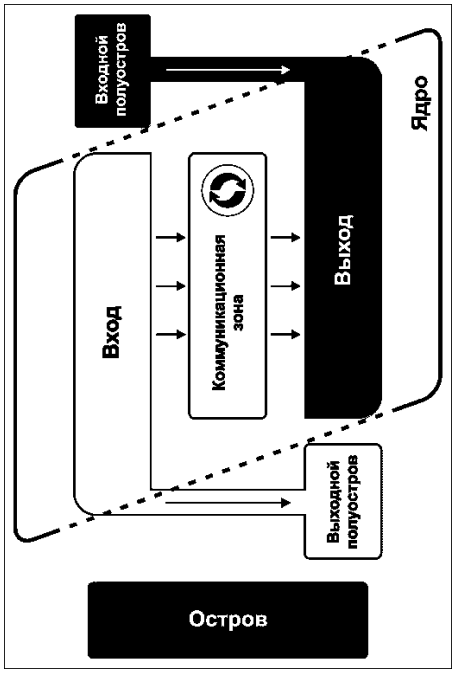


Рис. 3. Архитектура новостного Web

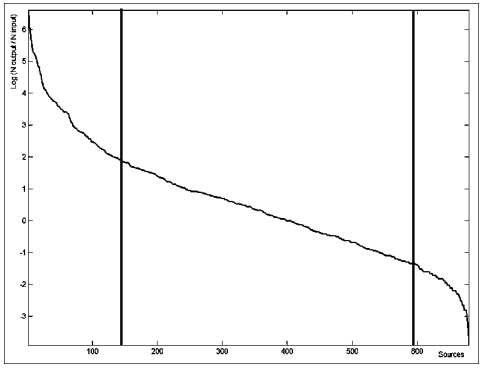


Рис. 4. Ранжированный график логарифма отношения количества ссылок

Интересным оказался график двумерного сечения значений $\log(N_{out} + 1)$, $\log(N_{in} + 1)$, где N_{out} - количество входящих ссылок, N_{in} - количество исходящих ссылок для каждого из источников (рис. 5). Этот график послужил основой идеальной схемы представления областей модели в зависимости от количества исходящих и входящих ссылок (рис. 6). В результате проведенных исследований была построена модель новостного Web-пространства, основанная на контекстных ссылках. Также предложены подходы к выявлению основных зон модели новостного Web-пространства и рассчитаны числовые соотношения различных зон модели.

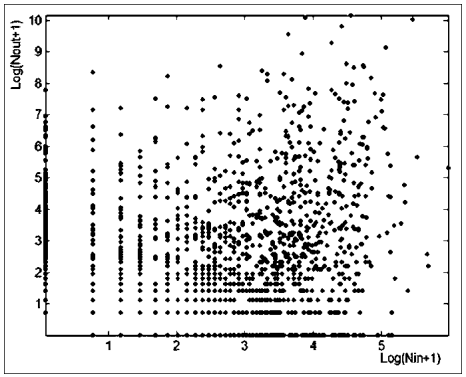


Рис. 5. График распределения зоны ядра в координатах «логарифм количества исходящих сообщений – логарифм количества входящих сообщений»

Вместе с тем данная модель предполагает дальнейшее совершенствование в следующих направлениях: более точной идентификации контекстных ссылок, совершенствовании 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).

Рис. 6. Представление областей модели в зависимости от количества исходящих и входящих ссылок

2. Математические модели информационных потоков

Изучение динамики, построение моделей информационных потоков [51], [54], [55], [6] является, бесспорно, важным и интересным, особенно учитывая то, что этот вопрос остается почти не исследованным [49], [50].

На протяжении последних десятилетий были достигнуты определенные успехи в решении проблемы старения информации в рамках модели Бартона-Кеблера [46], которая возникла в свое время из необходимости оценки реальных сроков использования научных работ, а также подходов Коула [48] и других авторов [3], [10], [28]. Со временем оказалось, что полученные результаты (а также подходы, которые лежали в их основе) могут быть полезными в более широком контексте проблем информационных технологий. Однако понимание процессов динамики информационных потоков требует более глубокого анализа и более совершенной техники.

В данной работе, в частности, предлагается рассмотрение динамики тематических потоков новостной информации в рамках логистической модели [2], [32]. Наряду с этим, выявлена ограниченность рассматриваемой модели, что открывает путь для дальнейших исследований.

Все Интернет-пространство можно с достаточной долей условности разделить на две составляющие - стабильную и динамическую [18], которые имеют очень разные характеристики с точки зрения интеграции информационных потоков. Стабильная составляющая Интернет содержит информацию "долгосрочного" плана, в то время как динамическая составляющая содержит постоянно обновляемые ресурсы. Некоторая часть этой составляющей со временем вливается в стабильную, однако большая часть "исчезает" из Интернет или попадает в сегмент "скрытого" Web-пространства, не доступного пользователям с помощью публичных информационно-поисковых систем.

Наиболее выраженным в плане динамики является, бесспорно, сегмент новостной информации. С одной стороны, он имеет высочайший уровень обновляемости, а с другой - в нем генерируются и распространяются на самом

деле большие объемы данных. Поэтому именно он выглядит наиболее подходящим для исследований. В частности, процессы старения информации, потери ее актуальности в известной модели Бартона-Кеблера [46] описываются уравнением, которое состоит из двух компонент:

$$m(t) = 1 - ae^{-T} - be^{-2T},$$

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

Вообще говоря, информационная динамика в сети обусловлена многими факторами, большинство из которых вообще не поддаются точному анализу. Однако в рамках задачи моделирования как разумное допущение можно предположить, что общий характер временной зависимости числа тематических публикаций в Сети определяется довольно простыми закономерностями, которые целиком допускают построение математических моделей.

В известных нам работах, посвященных изучению старения информации, используется модель Мальтуса [59] (возможно, с некоторыми модификациями, например, в виде суперпозиции двух кривых с разными параметрами в рамках приведенной выше модели Бартона-Кеблера). Преимуществом этой модели есть то, что уравнение Мальтуса имеет точное решение в виде очень простой и удобной функции - экспоненты, но с точки зрения интерпретации результатов она выглядит довольно сомнительной. Главной проблемой следует считать то, что экспонента есть монотонно возрастающая функция, а, следовательно, принципиально не может описывать процессы, которые по своей природе должны иметь локальные экстремумы.

То, что новости со временем теряют актуальность, и соответствующее количество публикаций уменьшается, не нуждается в доказательствах. Поэтому для получения более адекватной зависимости следует обратиться к более сложным моделям.

Одной из самых перспективных выглядит логистическая модель, которая была предложена П. Ферхлюстом [69] для описания динамики населения и

Р. Перлом [62] для биологических сообществ, а со временем хорошо зарекомендовала себя в целом ряде направлений научных исследований. Преимуществом этой модели есть в первую очередь то, что она объединяет относительную простоту формулирования задачи с возможностью варьировать решения с помощью набора параметров, которые могут иметь более или менее прозрачное физическое содержание.

Анализ информационных потоков, их моделирование сегодня становится одним из наиболее информативных методов количественного изучения динамики отдельных тематических направлений. По изменению величин информационных потоков судят о скорости развития как отдельных тематических направлений, так и всего информационного пространства.

Устойчивые статистические связи между отдельными сообщениями позволяют говорить о корреляции отдельных тематик, об эффективности ссылок на публикации предшественников, более ранние работы, цитирование, републикации и т.п.

Механизмы, которые базируются на обобщенных методах кластерного анализа разрешают обнаруживать сообщения в информационных потоках, которые формируют вокруг себя новые тематические направления. Кластерный анализ, теория фракталов и автомодельных процессов при их корректном применении разрешают количественно оценивать степень связи в тематических информационных потоках.

Из классической пространственно-векторной модели информационного пространства принято использовать модель TF*IDF, где *TF* – это локальная частота терма (Term Frequency), а *IDF* – величина, обратная частоте появления сообщений во всем информационном потоке, которые содержат этот терм (Inverse Document Frequency). В то время, как локальная частота терма в документе говорит о значимости терма в пределах документа, то обратная частота появления свидетельствует об уникальности терма во всем потоке документов. Поэтому произведение этих величин – достаточно удачный критерий определения значимости терма - веса. Предполагается, что новостные сообщения стареют,

теряя свою актуальность с интенсивностью, которая определяется некоторым эмпирическим законом. Для иллюстрации предположим, что это экспоненциальный закон (в дальнейшем будет показана корректность такого предположения для большого количества примеров). Один из предложенных подходов к такой части обобщения, как ранжирование сообщений, состоит в использовании параметрических множителей, которые зависят от времени, например, можно определить вес сообщения как произведение элементов типа $TF * IDF * e^{-\alpha t}$, где α - некоторая константа, t - интервал времени, которое прошло с момента появления сообщения в информационном потоке (значение α - это коэффициент полураспада актуальности сообщения, т.е., если предполагается применение экспонентной модели, это $e^{-\alpha t} = ½$, где t -период времени, которое определяется экспертным путем, на протяжении которого сообщение в результате старения теряет свою актуальность наполовину). Например, если предположить, что через сутки документ теряет половину своей актуальности, то имеем: $e^{-\alpha * 24} = ½$, и, соответственно, $\alpha = 0{,}025$.

Учет старения информации (потери части актуальности) имеет большое значение при аналитических исследованиях, создании информационных продуктов типа информационных портретов, основных сюжетов событий, ранжировании результатов работы информационно-поисковых систем. Даже приблизительная оценка скорости старения информации и отдельных документов имеет огромную практическую ценность, так как помогает держать в поле зрения только наиболее актуальную информацию.

С философской точки зрения понятие старения документов можно рассматривать как закономерный постоянный процесс уменьшения со временем их использования для получения необходимой пользователям информации, которая содержится в них. Процесс старения информации можно рассматривать как потерю информацией практической полезности для потребителя. Старение информации проявляется в том, что постоянно возникают новые документы, новые источники, которые содержат более полную, точную, достоверную информацию. Поэтому с целью экономии времени и ресурсов оправданно

первоочередное обращение именно к этим документам и источникам. При этом сложность использования закономерностей старения информационных сообщений состоит из разности характеристик уменьшения их использования во времени в разных предметных областях и для разных временных периодов. Степень старения информации неодинакова для документов разных видов и тематик. На скорость старения влияют в разной степени очень много факторов. Особенности старения информации органически связаны с тенденциями развития каждого тематического направления.

Для того, чтобы количественно оценить скорость старения информации, Р. Бартон и Р. Кеблер по аналогии с периодом полураспада радиоактивных веществ также ввели понятие «полупериода жизни» научных статей. Полупериод жизни в их понимании - это время, на протяжении которого была опубликована половина всех используемых в настоящее время документов относительно выбранного события или явления. Бартон и Кеблер определили периоды полураспада публикаций из физики (4,6 года), математики (10,5), геологии (11,8) [46].

2.1. Линейная модель информационных потоков

В некоторых случаях динамика тематических информационных потоков (повышения актуальности или старения информации) происходит линейно, то есть количество сообщений в момент времени t можно, соответственно, представить формулами:

$y(t) = y(t_0) + v(t - t_0)$,

$y(t) = y(t_0) – v(t - t_0)$,

где $y(t)$ – количество сообщений на время t, v – середняя скорость увеличения (уменьшения) интенсивности тематического информационного потока во времени (например, в результате старения).

Ниже (рис. 7 та 8) приведены примеры линейного роста количества сообщений из информационного потока в системе контент-мониторинга InfoStream, в которых встречаются слова, начинающиеся с шаблонов

«семантическ*» и «масон», соответственно. Из графиков, которые отображают динамику изменений понятий на протяжении года, сгруппированную по неделям, можно видеть, что уровень роста в первом случае больший (речь идет о постепенном росте популярности семантических сетей), а во втором случае - рост медленный, - это связано, прежде всего, с ростом количества источников, которые сканируются на протяжении года.

Содержательная составляющая информационного потока может быть количественно оценена как флюктуация информационного потока – изменение стандартного отклонения $\sigma(t)$, которое вычисляется по формулам:

$$\sigma(t_i) = \sqrt{\frac{1}{i}\sum_{k=0}^{i}\{y(t_k) - (y(t_0) + v(t_i - t_0))\}^2},$$

$$\sigma(t_i) = \sqrt{\frac{1}{i}\sum_{k=0}^{i}\{y(t_k) - (y(t_0) - v(t_i - t_0))\}^2}.$$

Как показано в работе [13], если эти величины изменяются как корень квадратный из времени, то процесс изменения публикаций по теме можно считать процессом с независимыми приращениями. При этом связями с предыдущими публикациями можно пренебречь.

В случае поведения стандартного отклонения по времени как $\sigma(t) \propto t^{\mu}$, чем большее значение μ, тем выше корреляция между текущими и предыдущими публикациями. В этих случаях μ характеризует степень связи между случайными событиями и принимает значения от ½ до 1.

2.2. Экспоненциальная модель информационных потоков

В некоторых случаях процесс увеличения (роста) актуальности или старения информации описывается экспоненциальной зависимостью, которую можно аппроксимировать такой формулой:

$N(t) = N(t_0)e^{\lambda(t - to)}$,

где λ - среднее относительное изменение интенсивности информационного потока.

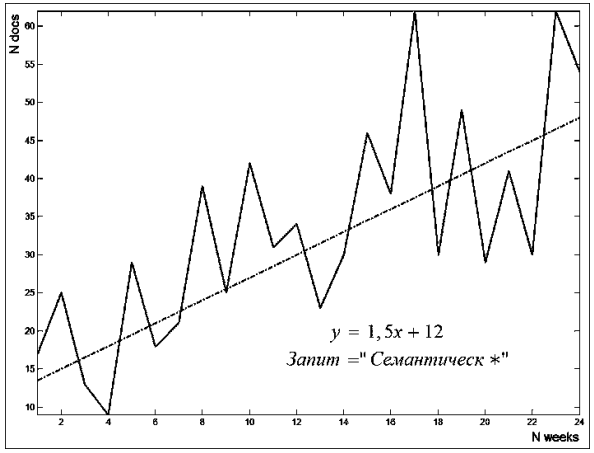


Рис. 7. Динамика количества откликов на запрос «семантическ»*

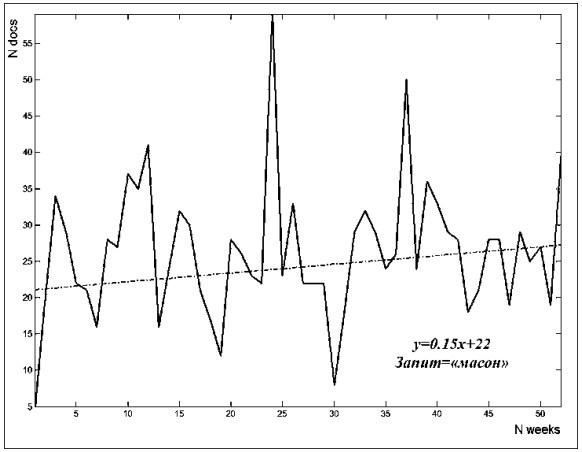


Рис. 8. Динамика появления слова «масон» в информационном потоке

Ниже (рис. 9 и 10) приведен пример экспоненциального роста количества сообщений из информационного потока системы контент-мониторинга InfoStream, в которых встретилось слово «блог». Экспоненциальный рост на протяжении 26 месяцев обусловлен ростом популярности нового средства общения в Интернет - «живых журналов».

Относительное изменение интенсивности в определенный момент времени вычисляется по формуле:

$\lambda(t_i)=(N(t_i) - N(t_{i-1}))/N(t_{i-1})$.

Изменение флюктуаций величины $\lambda(t_i)$ относительно среднего значения может быть оценено формулой:

$$\sigma(t_i) = \sqrt{\frac{1}{i}\sum_{k=0}^{i}\{\lambda(t_k) - \lambda\}^2}\,.$$

В этом случае также, если $\sigma(t)$ изменяется как корень квадратный от времени, то можно говорить о процессе с независимыми приращениями [13], корреляция между отдельными сообщениями несущественна. В случае наличия значительного количества зависимых сообщений справедливо: $\sigma(t) \propto t^{\mu}$, причем μ превышает ½, но ограничено 1.

Значение μ, которое превышает ½, говорит о наличии долгосрочной памяти системы. Такие системы порождают класс процессов, который получил название автомодельных, для которых предполагается корреляция между количеством сообщений информационных потоков в разные моменты времени.

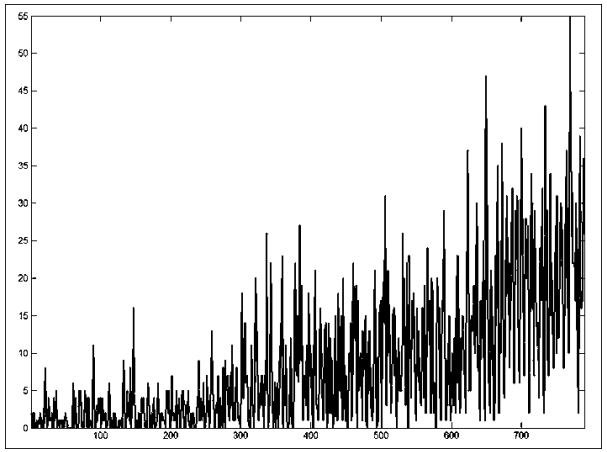


Рис. 9. Посуточный график появления термина «блог»

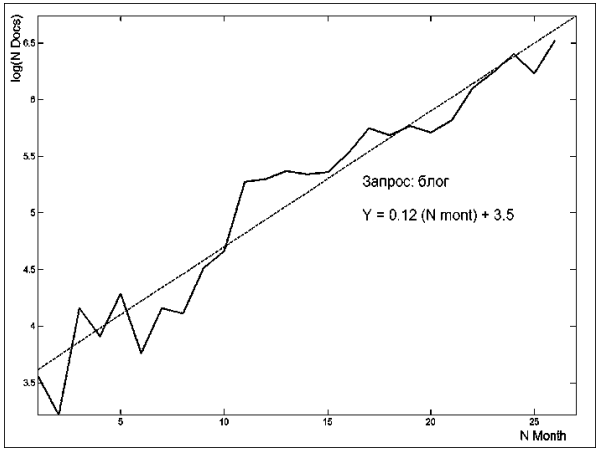


Рис. 10. Помесячный график появления термина «блог» в полулогарифмической шкале

Изучение флюктуаций информационных потоков показывает наличие статистической временной корреляции как на коротких, так и на продолжительных временных интервалах.

Новые надежды дает применение теории фракталов, которая позволяет говорить о проявлении свойств подобия для коммуникационных процессов на разных уровнях. Такой подход разрешил расширить представление об основных закономерностях коммуникационных процессов (в том числе и процессов роста актуальности или старения информации).

2.3. Логистическая модель информационных потоков

Рассмотрим общую картину динамики тематических информационных потоков, ограничившись механизмами, типичными для новостного сегмента Интернет.

Мы исходим из того, что организации-генераторы новостной информации в абсолютном большинстве работают в стационарном режиме, который может характеризоваться максимальной емкостью информационного пространства N (укажем, что вопрос о размерности параметров, а также об их измерении мы в данной публикации не рассматриваем). Это означает, что каждая организация-

генератор производит поток информации, в среднем постоянный по количеству как знаков, так и сообщений. Изменяются во времени лишь объемы сообщений, которые соответствуют той или другой теме. Другими словами, рост количества публикаций по одной теме сопровождается уменьшением публикаций на другие темы, так что для каждого промежутка времени T имеем:

$$\int_0^T \sum_{i=1}^{M} n_i(t)dt = NT, \qquad (2.1)$$

где $n_i(t)$ – количество публикаций в единицу времени, а M – общее количество всех возможных тем. Конечно, предполагается, что часть $n_i(t)$ всегда равняется нулю.

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

Следует отметить, что когда мы говорим о теме относительно информационного потока, то эти слова не следует воспринимать в рамках модели буквально. Под “темой” мы понимаем определенную абстракцию, связанную с активностью информационных источников. Конечно, она имеет связь с событиями в реальном мире, но субъективное оформление ее может оказаться не таким простым, как кажется на первый взгляд. Например, запуск нового космического корабля на Марс может вызвать поток публикаций о целесообразности перераспределения бюджета в пользу научных исследований.

Поэтому установить прямую связь между повышением активности источников-генераторов и ситуацией в окружающей обстановке возможно далеко не всегда. Здесь мы будем говорить о возникновении новой темы, принимая во внимание комплекс факторов, которые предопределяют рост количества публикаций в единицу времени. Локализация отдельной темы в семантическом пространстве и артикуляция ее в коммуникативных механизмах представляют отдельную проблему, которую мы не обсуждаем в рамках предлагаемой работы. Ограничимся лишь констатацией того, что в принципе она может быть решена в довольно широком спектре случаев. Главное для нас то, что темы возникают в

определенный момент времени и так же в определенный момент времени исчезают (т.е. теряют актуальность и перестают интересовать публику).

Теоретически можно предположить, что множества публикаций, ассоциированных с определенным набором тем, пересекаются, то есть существуют публикации, которые могут быть отнесены одновременно к нескольким разным темам. Вообще говоря, такая "политематичность" действительно является эффектом, с которым надо считаться, но мы в первом приближении будем считать, что его вклад не искажает общую картину.

Дальше, будем считать, что на протяжении времени своего существования (актуальности) тема фиксирует комплекс механизмов, которые приводят к росту количества публикаций, имеющих определенные общие черты. Разные темпы могут порождать разные по объему потоки публикаций, ведь в этом плане они не являются равнозначными. Поэтому на формальном уровне сопоставим с темой как абстрактным понятием два параметра: продолжительность (характерное "время жизни") λ и интенсивность D. В рамках данной работы мы будем считать интенсивность величиной постоянной. Это, конечно, упрощенный взгляд, но вполне достаточный для выяснения общих тенденций.

Продолжительность, как вытекает из сказанного выше, не обязательно должна совпадать с началом и окончанием какого-то события в реальном мире (или ряда событий). Она характеризует лишь характерный промежуток времени, на протяжении которого тема имеет оконченную актуальность. Интенсивность можем определить как величину, которая характеризует порожденное соответствующей темой количество публикаций, усредненное по промежутку λ.

Реакция медийных средств, описываемая величиной D, никогда не бывает мгновенной: всегда существует определенная задержка во времени. Чтобы учесть этот аспект, введем фактор опоздания τ.

Учитывая сказанное, мы можем предложить следующую качественную картину динамики тематических информационных потоков. Генерация информационных потоков имеет две основные составляющие: фоновую и собственно тематическую. Фоновая составляющая определяется наложением

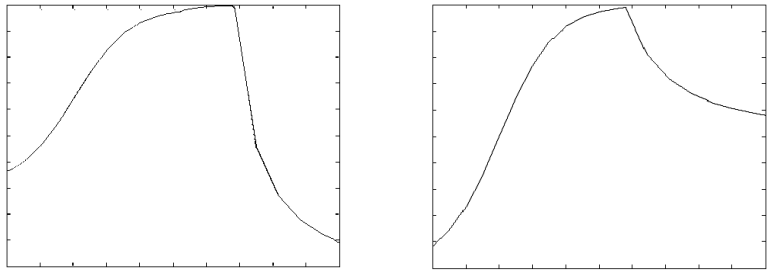
многих слабо связанных между собой факторов и при определенных условиях может приближаться (с точки зрения тематических распределений) к шуму. Но она обеспечивает публикацию более или менее стабильного количества материалов по принципу "Надо же что-то публиковать!"

Возникновение новой темы вызывает процесс (точнее говоря, комплекс процессов) перераспределения сетевых ресурсов в связи с появлением актуальных сюжетов. Объем фоновых публикаций снижается, а тематических – возрастает. Если продолжительности двух или более тем пересекаются, то соответствующие тематические публикации также начинают перераспределяться между ними, причем характер перераспределения определяется значениями параметров λ и D каждой темы. Когда же тема теряет актуальность, ассоциированные с ней ресурсы начинают переходить или в фоновые потоки, или в другие тематические.

В данной работе мы будем рассматривать именно тематическую составляющую, при чем сосредоточим внимание на динамике потоков, порожденных одной темой. Изучение взаимодействия нескольких тем представляет отдельное исследование, которое выходит за пределы поставленной нами задачи.

Приведем примеры лишь двух реальных информационных потоков, поведение которых попробуем учесть в модели, которую опишем ниже. В первом случае (рис. 11а) рассматривались публикации, которые сканировались системой мониторинга новостей из Интернет по тематике болезни и отхода от деятельности известного политического деятеля. К моменту обострения болезни объемы публикаций относительно его деятельности были на довольно высоком уровне. Болезнь значительно повысила количество публикаций, которое достигло верхнего уровня насыщения. Сведения относительно отхода от деятельности снизили количество публикаций до нижней планки, на этом уровне и осуществилась окончательная стабилизация. Второй пример - сенсационное избрание мэра большого города (рис. 11б). До избирательной кампании об этом лице было не слишком много публикаций в Интернет, что соответствовало нижней стабильной планке. Выбор и утверждение мэра сопровождались

значительным количеством публикаций как положительного, так и отрицательного характера (верхняя планка). Процесс дальнейшей деятельности мэра сопровождается объемами публикаций, которые соответствуют среднему стабилизационному уровню.



а) б)

Рис. 11. Примеры информационных потоков

При желании логистическую модель можно рассматривать как обобщение модели Мальтуса, которая, как известно, предусматривает пропорциональность скорости роста функции ее значению в каждый момент времени:

$$\frac{dn(t)}{dt} = kn(t), \qquad (2.2)$$

где k – некоторый коэффициент пропорциональности. Поскольку рассматривается динамика отдельного тематического потока, то далее не будем писать для величин $n_i(t)$ индексы, которые определяют тему.

Идея заключается в том, чтобы сделать коэффициент в уравнении Мальтуса функцией времени, причем так, чтобы решение не превышало заданного порогового значения. Существуют разные способы сделать это, но наиболее распространенным является использования константы, которая в явном виде ограничивает рост решения. В нашем случае с этой целью используем емкость N. Тогда правую часть выражения (2.1) можно представить в виде:

$$k(N - rn(t)), \qquad (2.3)$$

где k – коэффициент Мальтуса, а r – фактор, который описывает отрицательные для данной системы процессы, связанные с внутренними факторами.

Теперь нам надо учесть в явном виде параметры, которые характеризуют влияние темы на динамику публикаций.

Поскольку интенсивность D определена нами как константа, ее взнос представим следующим образом:

$$y(t) = \begin{cases} D, 0 < t \le \lambda \\ 0, t < 0, t > \lambda \end{cases} \qquad (2.4)$$

Соответственно, будем рассматривать отдельно две временные области: $0 < t \le \lambda$ с $D > 0$ и $t > \lambda$ с $D = 0$, для которых решениями являются функции $u(t)$ и $v(t)$. Полное решение получим путем “сшивки” на границе в точке λ:

$$n(t) = \begin{cases} u(t), 0 < t \le \lambda \\ v(t), t > \lambda \end{cases}$$

$$u(\lambda) = v(\lambda) \qquad (2.5)$$

Первой области соответствует процесс роста числа публикаций на данную тему в условиях ее ненулевой актуальности ($D > 0$) и, возможно, переход к состоянию насыщения, а второй – процесс сокращения числа публикаций, обусловленный потерей актуальности ($D = 0$).

Отнормировав параметры к пороговой величине N, представим уравнение для первой области в таком виде:

$$\frac{du(t-\tau)}{dt} = pu(t-\tau)(1 - qu(t-\tau)) + Du(t-\tau), \qquad (2.6)$$

$$u(0) = n_0$$

Величина p определяет нормированную вероятность в единицу времени появления публикации независимо от актуальности данной темы. Такой фактор отображает фоновые механизмы генерации информации (типичным примером может быть механическое перепечатывание материалов престижных информационных ресурсов). Величина D характеризует непосредственное влияние актуальности данной темы. Параметр q характеризует уменьшение

скорости роста количества публикаций и является величиной, обратной асимптотическому значению зависимости $u(t)$ при $D = 0$.

Начальное условие в (2.6) отражает два аспекта информационной динамики: во-первых, наличие фоновой составляющей информационных потоков, а во-вторых, неопределенность точного момента, когда определенная тема начинает вносить свой вклад в общий процесс генерации публикаций. Ввиду этого, в момент времени $t = 0$ существует некоторое количество публикаций, которые могут быть ассоциированы с данной темой.

Для второй области, соответственно, имеем:

$$\frac{dv(t-\lambda)}{dt} = pv(t-\lambda)(1-qv(t-\lambda)), \qquad v(\lambda) = u(\lambda) \tag{2.7}$$

Так как во второй области тема уже не оказывает влияния на динамику публикаций (она описывает инерционные по отношению к теме процессы), в уравнении (2.6) не включается фактор запаздывания τ. Предельное условие в (2.7) обеспечивает “сшивку” функций *u(t)* и *v(t)*.

Решение (2.6) имеет такой вид:

$$u(t) = \frac{u_s}{1 + (\frac{u_s}{n_0} - 1)\exp[-(p+D)(t-\tau)]}, \tag{2.8}$$

где u_s – асимптотическое значение u, величина которого определяет область насыщения (если, конечно, данная зависимость успевает ее достичь):

$$u_s = \frac{p+D}{pq}. \tag{2.9}$$

Заметим, что выражение (2.9) не зависит от значения n_0, что свидетельствует о несущественности для состояния насыщения информационной динамики начальных условий. Каким бы ни было начальное количество публикаций, насыщение будет определяться исключительно параметрами, которые характеризуют фоновую скорость роста числа публикаций, количественную меру актуальности и отрицательные для процесса факторы. А

потому с практической точки зрения можем пренебречь фоновыми факторами, которые плохо подвергаются изучению.

Кривая (2.8) имеет точку перегиба:

$$t_{\inf} = \frac{1}{p+D}\ln(\frac{u_s}{n_0} - 1) + \tau. \tag{2.10}$$

Таким образом, для первой области имеем так называемую S-подобную зависимость, а при $t \sim t_{\inf}$ зависимость (2.8) приближается к линейной и соответствует линейной модели.

Представим теперь для удобства (2.8) в другом виде:

$$\frac{u_s \exp[(p+D)(t-\tau)]}{\exp[(p+D)(t-\tau)] + (\frac{u_s}{n_0} - 1)} = \frac{u_s \exp[(p+D)t]}{\exp[(p+D)t] + (\frac{u_s}{n_0} - 1)\exp[(p+D)\tau]}. \tag{2.11}$$

Из этой записи видно, что при условии

$$t < \frac{1}{p+D}\ln(\frac{u_s}{n_0} - 1) + \tau = t_{\inf}. \tag{2.12}$$

зависимость $u(t)$ имеет экспоненциальный характер, причем ее выразительность определяется величиной опоздания τ. Итак, для значений t, значительно меньше $t_{\inf}$, наша модель совпадает с экспоненциальной моделью.

Типичная зависимость представлена на рис. 12.

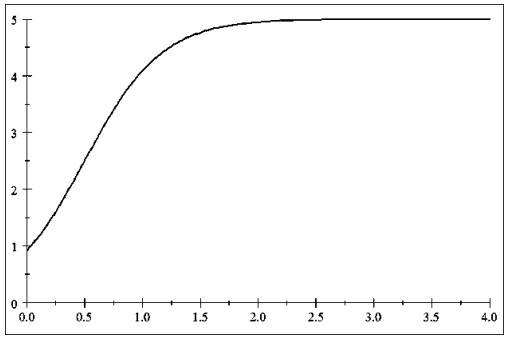


Рис. 12. Область роста

Перейдем ко второй области. Для нее решение имеет такой вид:

$$v(t) = \frac{u(\lambda)}{qu(\lambda) + (1 - qu(\lambda))\exp[-p(t-\lambda)]}. \qquad (2.13)$$

Если зависимость *u(t)* успевает достичь насыщения за промежуток времени $t < \lambda$, можем упростить решение (2.13), представив его следующим образом:

$$v(t) = \frac{v_s(p+D)}{p + D(1 - \exp[-p(t-\lambda)])}, \qquad (2.14)$$

где $v_s = 1/q$ асимптотическое значение зависимости *v(t)*.

Как и следовало ожидать, величина v_s также не зависит ни от начального условия, ни от условия “сшивки” на границе областей.

Во второй области динамика публикаций в первом приближении имеет экспоненциальный характер, что совпадает с результатами исследований.

Типичная зависимость для второй области представлена на рис. 13.

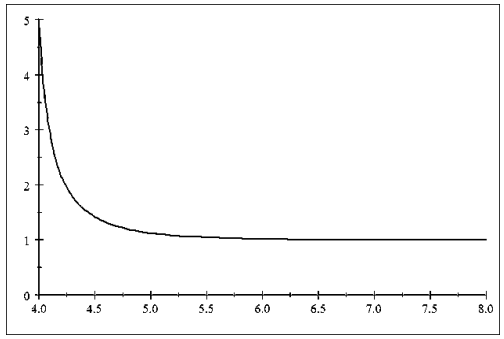


Рис. 13. Область спада

Итак, мы видим, что рассматриваемая зависимость имеет область насыщения u_s (при $t \leq \lambda$) и асимптотику v_s, которая описывает постепенное уменьшение числа публикаций к фоновому уровню. А это означает, что она, по крайней мере качественно, согласована с общими представлениями о характере информационной динамики, полученными на основе экспериментальных данных.

Кроме того, она неплохо совпадает с линейной и экспоненциальной моделями на определенных участках *t*.

Типичная полная зависимость *n*(*t*) приведена на рис. 14.

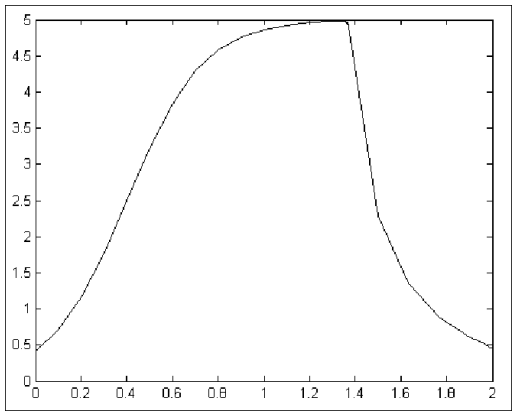


Рис. 14. Обобщенный график динамики тематического потока

Итак, предлагаемая модель правильно описывает (по крайней мере на уровне качественных свойств) временную зависимость плотности публикаций, порожденных отдельной темой. В частности, она содержит область насыщения, которую невозможно объяснить в рамках экспоненциальной модели.

Мы также видим, что полученная зависимость не является симметричной и имеет характерный “гребешок” на границе выделенных двух областей. Решения нашего уравнения для второй области, в отличие от первой, не имеет состояния насыщения: оно описывает близкий к экспоненциальному спад, который асимптотически приближается к нулю.

Такая интересная особенность поведения кривой на самом деле наблюдается на практике в определенной части случаев, но не во всех. Экспериментальные данные свидетельствуют о наличии еще двух типов зависимостей, которые не обсуждаются в данной публикации. Укажем лишь, что была рассмотрена простейшая реализация модели. Не исключено, что более

сложные ее модификации дадут возможность описать все основные разновидности реальной динамики.

Отдельную проблему информационной динамики представляют циклические процессы роста и снижения активности информационных ресурсов, не связанные с собственно информационными факторами (например, периодическое снижение количества публикаций по выходным дням).

Открытой остается проблема определения соотношения решений приведенных логистических уравнений с условием баланса тем (2.1).

Вместе с тем имеются веские основания для утверждения, что логистическая модель в самом деле описывает динамику определенной категории тематических информационных потоков.

2.4. Подход к анализу новостных потоков как дискретных сигналов

Одна из идей при исследовании новостной составляющей информационного пространства Интернет, к которой все чаще обращаются в настоящее время, заключается в анализе текстовых массивов как дискретных сигналов, определяемых частотно-семантическими рангами [49] ключевых слов или отдельных сообщений.

Рассмотрим модель, в которой аналогами дискретных сигналов выступают ключевые слова (наиболее ранговые термины) из сообщений или отдельные сообщения информационных потоков, порождаемых информационными Web-сайтами. В соответствии с приведенным ниже алгоритмом, каждому сообщению приписывается вес, который равен усредненной частоте появления во всем информационном потоке входящих в это сообщение значимых ключевых слов. Очевидно, чем меньше этот вес, тем документ более уникален.

Понятно, что для информационного наполнения модели необходимо использовать достаточно мощный текстовый корпус, который был доступен авторам - это база данных системы контент-мониторинга InfoStream [20].

Ниже приведен двухпроходный алгоритм формирования словаря уникальных слов из входного массива из N сообщений, а затем вычисления весов отдельных сообщений:

Этап 1: первичная обработка входного информационного массива

```
while количество необработанных сообщений из массива > 0 do
    чтение текущего сообщения
    for каждого сообщения do
        while не исчерпался список ключевых слов do
            for каждого ключевого слова do
                if ключевое слово уже входит в словарь
                then вес ключевого слова = вес ключевого слова + 1
                else добавить ключевое слово в словарь с весом 1
            end for
        end while
    end for
end while
```

Этап 2: повторная обработка информационного массива:

```
while количество необработанных сообщений из массива > 0 do
    чтение текущего сообщения
    вес сообщения = 0
    for каждого сообщения do
        счетчик ключевых слов = 0
        while не исчерпался список ключевых слов do
            for каждого ключевого слова do
                определение веса из словаря уникальных слов
                вес сообщения = вес сообщения + вес слова
                счетчик ключевых слов = счетчик ключевых слов + 1
            end for
        end while
    end for
    вес сообщения = вес сообщения / число ключевых слов
end while
```

Таким образом, вес сообщения определяется по формуле:

$$W_D = \frac{\sum_{w \in D} w}{|D|}$$

где W_D – вес сообщения, w – вес ключевого слова из сообщения, $|D|$ - количество ключевых слов в документе (в рассматриваемой модели $1 \le |D| \le 12$). Как видно, при значениях β в указанном выше диапазоне w является монотонно возрастающей функцией от n.

Как следует из алгоритма, каждое сообщение в данной модели рассматривается как массив ключевых слов (Bag of Words [65]), хотя при построении модели учитывались структурные особенности сообщений [41], в частности, при определении веса ключевых слов учет их местоположения в тексте.

В классической пространственно-векторной модели [66] значения рангов отдельных ключевых слов определяются формулой *TF* * *IDF*. В данном случае *TF* – это локальная частота ключевого слова (Term Frequency), а IDF – величина, обратная частоте встречаемости во всем потоке документов, содержащих данный терм (Inverse Document Frequency).

В то время как локальная частота ключевого слова в документе говорит об его значимости в пределах документа, то обратная частота встречаемости свидетельствует об уникальности ключевого слова во всем потоке документов.

В рассматриваемой модели в соотношении *TF*IDF* фактически анализируется лишь второй сомножитель (а точнее, обратная ему величина), исходя из того, что заведомо высокий уровень значений *TF* определяется процедурой выявления ключевых слов, выполняемой ранее системой контент-мониторинга.

В рамках модели в качестве веса ключевых слов используется частота их появлений во входном информационном потоке. В свою очередь, эта частота зависит от объема самого потока и от количества уникальных слов, т.е. объема автоматически формируемого словаря уникальных слов. В компьютерной лингвистике эмпирический закон Хипса [56] связывает объем документа с

объемом словаря уникальных слов, входящих в этот документ. В соответствии с законом Хипса, эти значения связываются соотношением:

$$v(n) = Kn^{\beta}$$

где v – это объем словаря уникальных слов, составленный из текста, состоящего из n уникальных слов. K и β – определяемые эмпирически параметры. Для европейских языков K принимает значения от 10 до100, а β - от 0.4 до 0.6.

В случае анализа не полных текстов, а фиксированного количества нормированных ключевых слов, эти параметры изменяются, однако сама закономерность Хипса остается в силе (рис. 15).

Джордж Зипф [71] экспериментально показал, что, если для какого-либо достаточно большого текста составить список всех встретившихся в нем слов, а затем ранжировать эти слова в порядке убывания частоты встречаемости в тексте, то для любого слова произведение его ранга в этом списке и частоты встречаемости в тексте будет величиной постоянной.

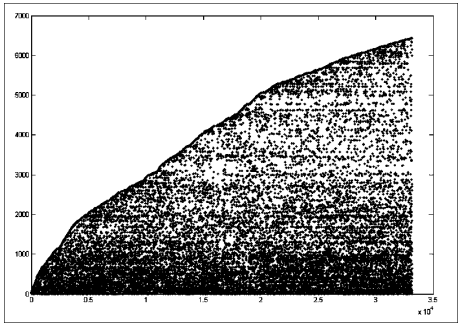


Рис. 15. График зависимости количества уникальных ключевых слов от общего количества ключевых слов потока подчиняется закону Хипса. При этом K = 4, β = 0,65

В рассматриваемой же нами модели в соответствии с приведенным выше алгоритмом распределение весов ключевых слов вполне вписывается в закон Зипфа (рис. 16), сформулированный изначально для ранговых распределений ненормированных слов в полнотекстовых документах. Однако в модели вместо ранжированного сортированного словаря используется простой порядковый номер. Феномен объясняется тем, что, в соответствии с положениями

математической статистики, большая часть наиболее часто встречающихся слов попадает в некоторое ограниченное количество первых по порядку сообщений.

Статистически связанная с названными выше закономерностями зависимость параметров распределения весов отдельных сообщений от их порядковых номеров в потоке (рис. 17) имеет вполне определенное смысловое объяснение. Оказывается, что амплитуда этого распределения возрастает с увеличением количества сообщений в потоке (рис. 18). Действительно, средний вес уникального ключевого слова равен общему числу слов из потока, разделенному на количество уникальных слов:

$w(n) = n/v(n) = n^{1-\beta} / K.$

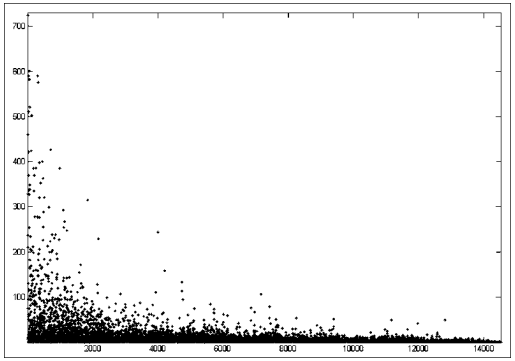


Рис. 16. Зависимость частоты уникальных слов в потоке от их порядковых номеров

Этому же значению равно и математическое ожидание веса отдельного сообщения из потока.

Изображенные на рис. 18. основные области графика дискретного сигнала, соответствующего информационному потоку, можно охарактеризовать следующим образом:

- горизонтальные зоны:

  1,2,3 – топ-новости; 4,5,6 – мэйнстрим; 7,8,9 – маргинальная зона;

- вертикальные зоны:

  1,4,7 – устаревающие сообщения; 2,5,8 –основная тематика; 3,6,9 - последние известия

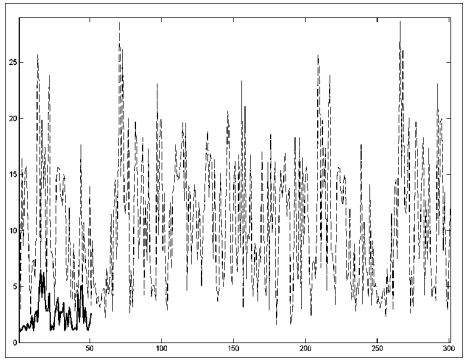


Рис. 17. Графики зависимости веса сообщений от их номеров в потоке. Рассматриваются два информационных потока (50 и 300 сообщений).

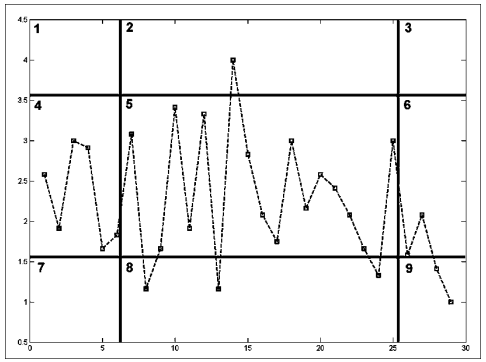
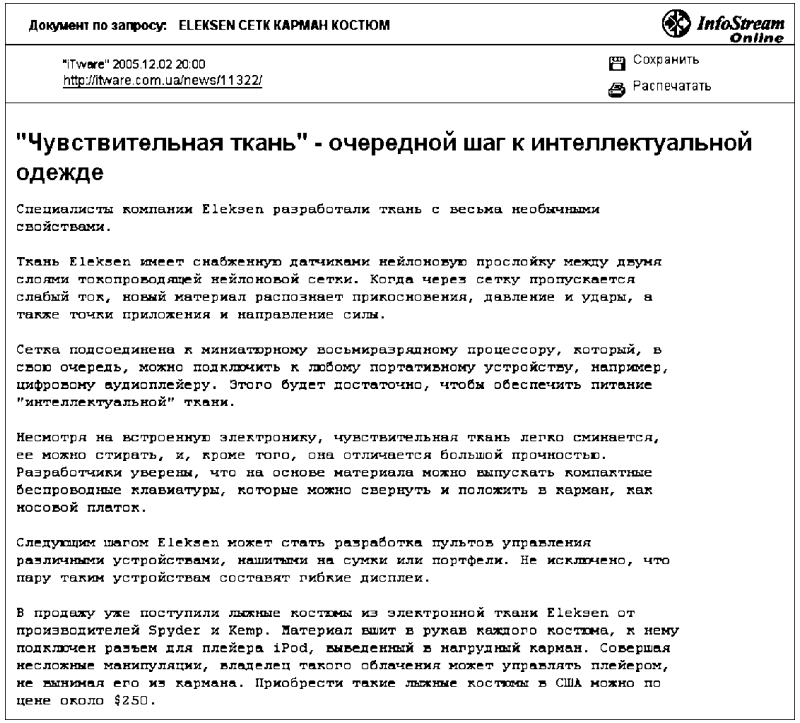


Рис. 18. Основные области графика распределения весов сообщений

На рис. 19 приведен документ, попавший в маргинальную зону при анализе потоков сообщений по компьютерной тематике, полученных с веб-сайта ITWARE (http://itware.com.ua). Этот пример с очевидностью подтверждает уникальность содержания сообщений из этой области по сравнению с мэйнстрим-сообщениями по информационным технологиям. Это всего лишь одно из многих практических подтверждений корректности данной модели, подхода к созданию инструментария в рамках системы контент-мониторинга, обеспечивающего просмотр маргинальных сообщений по тематике, определяемой запросом, т.е. фактически дающего ответ на вопрос, о чем пишут меньше всего в рамках данной тематики в последнее время. Этот инструментарий логически дополняет уже

существующий в системе InfoStream сервис получения сюжетов из наиболее популярных сообщений [18].

В заключение заметим, что предложенная модель охватывает лишь некоторые частотно-семантические подходы к рассмотрению текстовых информационных потоков как дискретных сигналов. Получены первые результаты исследования, которое может включать в себя более полный учет структурных особенностей текстов, анализ корреляции сигналов, фильтрацию типа «сигнал-шум» и т.д. Можно также предположить, что к обработке текстовых потоков будут применимы такие популярные сегодня техники обработки сигналов, как анализ главных компонент, слепое разделение источников, вэйвлеты.



Документ по запросу: ELEKSEN СЕТК КАРМАН КОСТЮМ

InfoStream Online

"iTware" 2005.12.02 20:00
http://itware.com.ua/news/11322/

Сохранить
Распечатать

"Чувствительная ткань" - очередной шаг к интеллектуальной одежде

Специалисты компании Eleksen разработали ткань с весьма необычными свойствами.

Ткань Eleksen имеет снабженную датчиками нейлоновую прослойку между двумя слоями токопроводящей нейлоновой сетки. Когда через сетку пропускается слабый ток, новый материал распознает прикосновения, давление и удары, а также точки приложения и направление силы.

Сетка подсоединена к миниатюрному восьмиразрядному процессору, который, в свою очередь, можно подключить к любому портативному устройству, например, цифровому аудиоплейеру. Этого будет достаточно, чтобы обеспечить питание "интеллектуальной" ткани.

Несмотря на встроенную электронику, чувствительная ткань легко сминается, ее можно стирать, и, кроме того, она отличается большой прочностью. Разработчики уверены, что на основе материала можно выпускать компактные беспроводные клавиатуры, которые можно свернуть и положить в карман, как носовой платок.

Следующим шагом Eleksen может стать разработка пультов управления различными устройствами, нашитыми на сумки или портфели. Не исключено, что пару таким устройствам составят гибкие дисплеи.

В продажу уже поступили лыжные костюмы из электронной ткани Eleksen от производителей Spyder и Kemp. Материал вшит в рукав каждого костюма, к нему подключен разъем для плейера iPod, выведенный в нагрудный карман. Совершая несложные манипуляции, владелец такого облачения может управлять плейером, не вынимая его из кармана. Приобрести такие лыжные костюмы в США можно по цене около $250.

Рис. 19. Сообщение по компьютерной тематике из маргинальной зоны

3. Фрактальные свойства информационного пространства

Термин *фрактал (*от латинского слова fractus – дробный), был предложен Бенуа Мандельбротом в 1975 году для обозначения нерегулярных самоподобных математических структур. Популярная сегодня фрактальная геометрия получила свое название лишь в 1977 году благодаря его книге «The Fractal Geometry of Nature». В работах Мандельброта использованы научные результаты других ученых, работавших в этой же области (прежде всего, Пуанкаре, Кантора, Хаусдорфа). Основное определение фрактала, данное Мандельбротом, звучало так: *"Фракталом называется структура, состоящая из частей, которые в каком-то смысле подобны целому"*.

В самом простом случае небольшая часть фрактала содержит информацию обо всем фрактале. Строгое определение самоподобных множеств было дано Дж. Хатчинсоном в 1981 году. Он назвал множество самоподобным, если оно состоит из нескольких компонент, подобных всему этому множеству, т.е. компонент, получаемых афинными преобразованиями - поворотом, сжатием и отражением исходного множества.

Однако самоподобие – это хотя и необходимое, но далеко не достаточное свойство фракталов. Ведь нельзя же, в самом деле, считать фракталом точку, или плоскость, расчерченную на клетки. Главная особенность фракталов заключается в том, что их размерность не укладывается в привычные геометрические представления. Фракталам характерна геометрическая «изрезанность». Поэтому используется специальное понятие фрактальной размерности, введенное Ф. Хаусдорфом и А. Безиковичем. Эта размерность не соответствует привычным для нас длине, площади или объему (размерности 1, 2 или 3, соответственно). Размерность фракталов не является целым числом, характерным для привычных геометрических объектов. Вместе с тем в большинстве случаев фракталы

напоминают объекты, плотно занимающие реальное пространство, но не использующее его полностью.

В реальной жизни фрактальные объекты имеют вполне определенные границы фрактальности, в том числе и самоподобия. Тем не менее фракталы – это очень удобная и наглядная абстракция, которая сегодня уже широко применяется при моделировании естественных процессов. При этом спектр применения фракталов постоянно расширяется, сегодня он применяется и к моделированию информационного пространства.

Один из лучших примеров проявления фракталов в природе – структура береговых линий. Действительно, на километровом отрезке побережье выглядит столь же изрезанным, как и на стокилометровом.

Опыт показывает, что длина береговой линии L зависит от масштаба l, в котором проводятся измерения, и увеличивается с уменьшением последнего по степенному закону $L = \Lambda\, l^{1-\alpha}$, $\Lambda = const$. Так, например, для побережья Великобритании (рис. 20) $\alpha \approx 1.24$, то есть, так называемая фрактальная размерность береговой линии Великобритании равна 1.24 .

Рис. 20. Береговая линия побережья Великобритании

В настоящее время информационное пространство в целом, ввиду его объемов и динамики изменения, принято рассматривать как стохастическое. Во

многих моделях информационного пространства изучаются структурные связи между тематическими множествами, входящими в это пространство. При этом численные характеристики этих множеств подчиняются гиперболическому закону (с возможными степенными поправками). Сегодня в моделировании информационного пространства все чаще используется фрактальный подход, базирующийся на свойстве самоподобия информационного пространства, т.е. сохранение внутренней структуры множеств при изменениях их размеров или масштабов их рассмотрения извне.

Самоподобие информационного пространства выражается прежде всего в том, что при его лавинообразном росте в последние десятилетия, частотные и ранговые распределения, получаемые в таких разрезах, как источники, авторы, тематика практически не меняют своей формы. Т.е. применение теории фракталов при анализе информационного пространства позволяет с общей позиции взглянуть на закономерности, составляющие основы информатики. Например, тематические информационные массивы сегодня представляют развивающиеся самоподобные структуры, которые по своей сути являются стохастическими фракталами, так как их самоподобие справедливо лишь на уровне математических ожиданий, например, распределения кластеров по размерам.

В информационном пространстве возникают, формируются, растут и размножаются кластеры – группы взаимосвязанных документов. Системы, основанные на кластерном анализе, самостоятельно выявляют новые признаки объектов и распределяют объекты по новым группам.

Чем же определяется природа фрактальной структуры информационного пространства, порождаемого такими кластерными структурами? С одной стороны, параметрами ранговых распределений, а с другой стороны, механизмом развития информационных кластеров, который отражает природу информационного пространства. Появление новых публикаций увеличивает размерность уже существующих кластеров и является причиной образования новых.

Фрактальные свойства характерны для кластеров информационных Web-сайтов, на которых публикуются документы, соответствующие определенным тематикам. Эти кластеры, как наборы тематических документов, представляют собой фрактальные структуры, обладающие рядом уникальных свойств. Например, российскими исследователями (С. Иванов и др.), определена фрактальная размерность подобных информационных массивов, изменяющаяся в пределах от 1.05 до 1.50, что свидетельствует о небольшой плотности заполнения кластеров документами по одной теме.

Как один из основных законов, отражающих самоподобие информационного пространства, можно назвать закон Зипфа. В 1949 году профессор филологии из Гарварда Дж. Зипф собрал достаточный статистический материал и экспериментально показал, что распределение слов естественного языка подчиняется закону: *"Если к какому-либо достаточно большому тексту составить список всех встретившихся в нем слов, а затем ранжировать эти слова, т.е. расположить их в порядке убывания частоты встречаемости в данном тексте и пронумеровать в возрастающем порядке, то для любого слова произведение его порядкового номера (ранга) в этом списке и частоты его встречаемости в тексте будет величиной постоянной."* Ученый описал обнаруженную им закономерность распределения слов в текстах на английском языке:

- небольшое количество слов, таких как "the", "and" в английском языке, которые имеют очень высокий ранг;
- среднее количество слов имеет средний ранг;
- большое количество слов имеет очень низкий ранг.

Таким образом: $f * r = c$, где f - частота встречаемости слова в тексте; r - ранг (порядковый номер) слова в списке; c - эмпирическая постоянная величина. Эту закономерность зависимости частоты от ранга называют первым законом Зипфа. То есть, зависимость количества слов с данной частотой от частоты - гипербола с постоянными параметрами для всех текстов в пределах одного языка. Значение константы в разных языках различно, но внутри одной языковой группы

остается неизменным. Так, например, для английских текстов константа Зипфа равна приблизительно 0,1. Для русского и украинского языков коэффициенты Зипфа составляют приблизительно 0,06-0,07.

Зипф сформулировал еще одну закономерность, близкую по смыслу к своему первому закону. Он определил, что частота и количество слов, входящих в текст с этой частотой, также взаимосвязаны. Если построить диаграмму, отложив по одной оси частоту вхождения слова, а по другой - количество слов, входящих в текст с данной частотой, то получившаяся кривая будет сохранять свои параметры для всех текстов в пределах одного языка. Однако на каком бы языке текст ни был написан, форма кривой Зипфа останется неизменной – могут отличаться лишь коэффициенты. Эта закономерность получила название второго закона Зипфа - "количество - частота".

Основатель теории фракталов Б. Мандельброт предложил теоретическое обоснование закона Зипфа, полагая, что можно сравнивать язык текста с кодированием. Исходя из требований минимальной стоимости сообщений, Мандельброт математическим путем пришел к аналогичной первому закону Зипфа зависимости $f*r^e = c$, где e - близкая к единице переменная величина, которая может изменяться в зависимости от свойств текста и языка. Постоянство коэффициента e сохраняется только в центральной зоне диаграммы распределения. По относительной величине той или иной зоны на графике можно судить о характеристиках рассматриваемой в тексте области знаний. Существуют также закономерности, открытые другими учеными (прежде всего, Брэдфордом - для периодических изданий и Лотки – для распределения авторов), являющиеся уточняющими следствиями закономерностей Зипфа, и также свидетельствующими о самоподобии информационного пространства.

Теория фракталов тесно связана с кластерным анализом, решающим задачу выделения компактных групп объектов с близкими свойствами. Кластеризация сегодня применяется при реферировании больших документальных массивов, определении взаимосвязанных групп документов, для упрощения процесса просмотра при поиске необходимой информации, нахождения уникальных

документов из коллекции, выявления дубликатов или близких по содержанию документов.

Фрактальный принцип самоподобия предполагает бесконечное дробление набора объектов с сохранением их свойств. В данном случае можно наблюдать подобие сюжетных цепочек, получаемых при уточнении запроса (конечно, в определенных рамках). Вместе с тем сегодня многими исследователями рассматривается не дробление, а естественный рост размеров информационного пространства.

Свойства самоподобия фрагментов информационного пространства наглядно демонстрирует новый интерфейс, представленный на Web-сайте службы News Is Free (http://newsisfree.com). На этом сайте отображается состояние информационного пространства в виде ссылок на источники и отдельные сообщения (рис. 21). При этом учитываются два основных параметра отображения – ранг популярности и «свежесть» информации. В рамках этой модели можно наблюдать «дробление» групп источников при увеличении ранга популярности и «свежести» изданий. Когда этот ранг становится достаточно высоким, дробление не позволяет без особых усилий читать названия источников и идентифицировать отдельные документы.

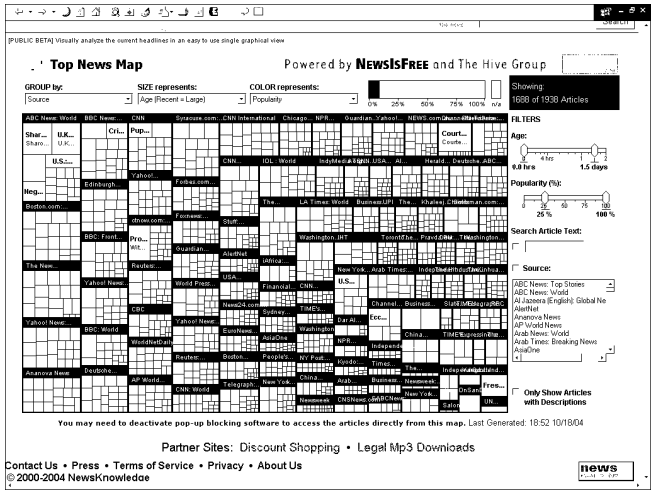


Рис. 21. Кластеры публикаций службы News Is Free

Пространство интернет-новостей, являясь, пожалуй, самой динамичной частью Web-пространства, характеризуется большим количеством контекстных и гипертекстовых ссылок, топология которых изложена в первой главе данной моногафии.

Как и в случае модели Бредера для Web-пространства, топология и характеристики модели новостной части Интернет оказались примерно одинаковыми для различных его подмножеств, подтверждая наблюдение о том, что "информационное пространство новостей - это фрактал", т.е. свойства всей структуры этого пространства верны и для его отдельных подмножеств.

С другой стороны, информационное пространство можно рассматривать как среду, в которой возникают и развиваются кластерные структуры, которые можно изучать и моделировать, используя как методическую основу теорию фракталов.

3.1. Фрактальные свойства тематических информационных потоков

Новостную составляющую информационного пространства Интернет можно рассматривать как мощный информационный поток [1], характеризующийся определенным набором параметров, среди которых выделяются такие, как источники информации (Web-сайт) и тематики. Именно их можно рассматривать как лежащие на поверхности основы для кластеризации [64].

В то время как для традиционных средств научной коммуникации подходы к кластеризации с точки зрения теории фракталов были впервые исследованы Ван Рааном, анализировавшим массивы статей и связи, образуемые цитированием, информационные потоки сообщений из Интернет до последнего времени не ассоциировались с фракталами, что связано с проблемами идентификации информационных потоков как фрактальных множеств, а также с трудностью нахождения основ для построения кластеров — сообщений в политематических потоках, порождающих многократное цитирование.

По этой же причине в рамках данной работы исследуются количественные характеристики лишь тематических информационных потоков, которые характеризуются итеративностью при формировании и вполне доступны как для количественного, так и для качественного анализа.

Объемы сообщений в тематических информационных потоках образуют временные ряды. Для исследования временных рядов сегодня все шире используется теория фракталов, традиционная область применения которой — фрактальная геометрия, обработка изображений и т.п. [18]. Вместе с тем временные ряды, порождаемые тематическими информационными потоками, также обладают фрактальными свойствами [35] и могут рассматриваться как стохастические фракталы [12], [13]. Этот подход расширяет область применения теории фракталов на информационные потоки, динамика которых описывается средствами теории случайных процессов.

С другой стороны, теория фракталов рассматривается как подход к статистическому исследованию, который позволяет получать важные характеристики информационных потоков, не вдаваясь в детальный анализ их внутренней структуры и связей. Одним из основных свойств фракталов является самоподобие (скейлинг). Как показано в работах С.А. Иванова, для последовательности сообщений тематических информационных потоков в соответствии со скейлинговым принципом, количество сообщений, резонансов на события реального мира пропорционально некоторой степени количества источников информации (кластеров) и итерационно продолжается в течение определенного времени. Точно так же, как и в традиционных научных коммуникациях, растущее множество сообщений в Интернет по одной тематике во времени представляет собой динамическую кластерную систему, возникающую в результате итерационных процессов. Этот процесс объясняется републикациями, прямой или совместной цитируемостью, различными публикациями — отражениями одних и тех же событий реального мира, прямыми ссылками и т.д. Кроме того, для большинства тематических информационных

потоков наблюдается увеличение их объемов, причем на коротких временных интервалах — линейный рост, а на длительных — экспоненциальный.

Фрактальная размерность в кластерной системе, соответствующей тематическим информационным потокам, показывает степень заполнения информационного пространства сообщений в течение определенного времени:

$$N_{публ} = \varepsilon^{\rho} N_k(t)^{\rho}, \qquad (3.1)$$

где $N_{публ}$ — размер кластерной системы (общее число электронных публикаций в информационном потоке); N_k — размер — число кластеров (тематик или источников); ρ — фрактальная размерность информационного массива; ε — коэффициент масштабирования. В приведенном соотношении между количеством сообщений и кластеров проявляется свойство сохранения внутренней структуры множества при изменении масштабов его внешнего рассмотрения.

По мнению С.А. Иванова, все основные законы научной коммуникации, такие как законы Парето, Лотки, Бредфорда, Зипфа, могут быть обобщены именно в рамках теории стохастических фракталов.

Сегодня, в связи с развитием теории стохастических фракталов, становится популярной такая характеристика временных рядов как показатель Херста (H). В книге Е. Федера [35] показано, что он связан с традиционной «клеточной» фрактальной размерностью (Θ) простым соотношением:

$$\Theta = 2 - H. \qquad (3.2)$$

Условие, при котором показатель Херста связан с фрактальной «клеточной» размерностью в соответствии с формулой (2), определено Е. Федером следующим образом: *«… рассматривают клетки, размеры которых малы по сравнению как с длительностью процесса, так и с диапазоном изменения функции; поэтому соотношение справедливо, когда структура кривой, описывающая фрактальную функцию, исследуется с высоким разрешением, т.е. в локальном пределе».* Еще одним важным условием является самоаффинность функции. Не вдаваясь в подробности заметим, что для информационных потоков это свойство интерпретируется как самоподобие, возникающее в результате процессов их

формирования. Можно отметить, что указанными свойствами обладают не все информационные потоки, а лишь те, которые характеризуются достаточной мощностью и итеративностью при формировании. При этом временные ряды, построенные на основании мощных тематических информационных потоков, вполне удовлетворяют этому условию. Поэтому при расчете показателя Херста фактически определяется и такой показатель тематического информационного потока как фрактальная размерность.

Известно, что показатель Херста представляет собой меру персистентности — склонности процесса к трендам (в отличие от обычного броуновского движения). Значение $H > ½$ означает, что направленная в определенную сторону динамика процесса в прошлом, вероятнее всего, повлечет продолжение движения в том же направлении. Если $H < ½$, то прогнозируется, что процесс изменит направленность. $H = ½$ означает неопределенность — броуновское движение.

Для изучения фрактальных характеристик тематических информационных потоков изучались значения показателя Херста за определенный период для временных рядов, составленных из количества относящихся к ним сообщений. Показатель Херста связывают с коэффициентом нормированного размаха (R/S), где R — вычисляемый определенным образом «размах» соответствующего временного ряда, а S — стандартное отклонение.

Показатель Херста вычисляется по следующему алгоритму. Сначала вычисляется среднее значение измеряемой переменной (в нашем случае количество сообщений в информационном потоке) за N дней:

$$< \xi >_N = \frac{1}{N} \sum_{t=1}^{N} \xi(t) . \qquad (3.3)$$

Затем рассчитывается накопившееся отклонение ряда измерений $\xi(t)$ от среднего $< \xi >_N$:

$$X(t,N) = \sum_{u=1}^{t} (\xi(u) - < \xi >_N) . \qquad (3.4)$$

После этого определяется разность максимального и минимального накопившегося отклонения, которая и называется «размахом»:

$$R(N) = \max_{1 \le t \le N} X(t,N) - \min_{1 \le t \le N} X(t,N) . \qquad (3.5)$$

Стандартное отклонение рассчитывается по известной формуле:

$$S = (\frac{1}{N} \sum_{t=1}^{N} (\xi(t) - < \xi >_N)^2)^{1/2} . \qquad (3.6)$$

В свое время Херст экспериментально обнаружил, что для многих временных рядов справедливо:

$$R / S = (N / 2)^H . \qquad (3.7)$$

Именно коэффициент H и получил название показателя Херста.

В качестве экспериментальной базы для исследования фрактальных свойств тематических информационных потоков использовалась система контент-мониторинга InfoStream.

Тематика исследуемого информационного потока определялась запросом к системе InfoStream, состоящим всего из одного слова «Microsoft». Ретроспективный период исследования составлял весь 2005 год и 2 месяца 2006 года, т.е. 424 дня ($N = 424$). В результате поиска было найдено 42357 релевантных документов.

Исходные данные были получены из интерфейса режима «Динамика появления понятий» (рис. 22). На основании обработки этих данных была получена полная картина экспериментальных данных — временной ряд за указанный период (рис. 23).

Для этого временного ряда по формуле (3.6) было вычислено стандартное отклонение ($S = 43{,}71$). Одновременно, с помощью механизма формирования основных сюжетов, входящего в состав системы InfoStream, были определены основные события, приведшие к возникновению пиковых значений на диаграмме.

На рис. 24 представлена динамика накопления отклонения, которая была вычислена в соответствии с формулой (3.4) и позволила в соответствии с формулой (3.5) определить «размах» этого параметра ($R = 1207{,}64$).

И наконец, для значения $N = 424$ по формуле (3.7) был вычислен показатель Херста, который оказался равным 0,62, что свидетельствует о положительной персистентности всего временного ряда.

Рис. 22. Фрагмент диаграммы динамики встречаемости понятия «Microsoft»

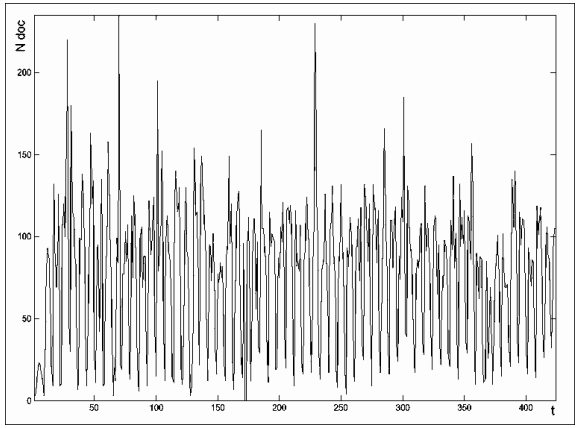


Рис. 23. Временной ряд встречаемости понятия за весь период. Пиковые значения: встречи в Давосе (конец января 2005 г.); признание журналом Forbes Б. Гейтса самым богатым человеком в мире (март 2005 г.); публикация журналом Time 100 самых влиятельных людей планеты (апрель 2005 г.); атака сетевого червя ZOTOB (август 2005 г.); 50-летний юбилей Б. Гейтса (конец октября 2005 г.)

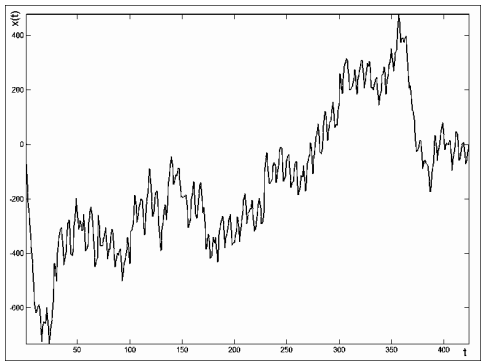


Рис. 24. Динамика накопления отклонения

Кроме того, были выполнены расчеты показателей Херста для всех значений *N*, начиная с 5, результаты которых приведены на рис. 25.

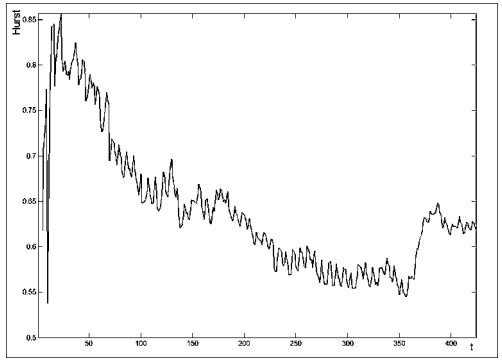


Рис. 25. Значения показателя Херста для различных временных интервалов

Изучение такой характеристики, как показатель Херста позволяет прогнозировать динамику информационных потоков, сообщения которых отражают процессы, происходящие в реальном мире.

Приведенные в примере данные подтвердили лежащее в основе исследования предположение об итеративности процессов в информационном пространстве. Републикации, цитирование, прямые ссылки и т.п. порождают самоподобие, проявляющееся в устойчивых статистических распределениях и

известных эмпирических законах. Скейлинговый принцип объясняется также сходством ментальности авторов, публикующих сообщения в Интернет. Вместе с тем различные маркетинговые, рекламные, PR-кампании ведут к скачкообразным изменениям в стабильных статистических закономерностях, резким скачкам и искажениям по сравнению со стандартными статистическими распределениями.

В результате эксперимента также подтверждено наличие статистической корреляции в информационных потоках на длительных временных интервалах.

В частности, на рассматриваемом примере показана персистентность процесса, что говорит об общем среднем увеличении публикаций о компании Microsoft, периодическом появлении пиков, связанных, как правило, с двумя подтемами-кластерами — личностью Билла Гейтса (четыре из пяти топ-кластеров) и отражениями вирусных атак (пятый топ-кластер).

Естественно, описанные результаты исследований могут использоваться не только для приведенного тематического информационного канала. Своего исследования ждут кластеры, порождаемые в соответствии и с другими принципами, например, близкими по направлениям источниками информации (Web-сайтами, сетевыми СМИ, блогами и др.)

3.2. Стабильность источников информации

Один из возможных подходов к решению проблемы изучения сетевого информационного пространства основан на представлении его некоторым множеством источников, порождающих информационные потоки. Предполагается, что динамика этих потоков в определенном смысле более содержательна, чем динамика составляющих их сообщений.

При этом можно отметить разнообразный диапазон паметров этих источников, как по объемам публикуемой информации, так и по содержанию – от сообщений серьезных информационных агентств – до «живых журналов» школьников.

Источники информации, очевидно, характеризуются уровнем стабильности. Примером стабильных источников могут служить крупные информационные

агентства, регулярно поставляющие потребителям примерно одинаковые объемы информации на протяжении длительного времени, а примером нестабильных – «живые журналы», многие из которых активно действуют в течение нескольких дней, а затем угасают.

Нестабильные источники по-своему интересны хотя бы тем, что, видимо, именно они ответственны за хаотичность динамической части сетевого информационного пространства. Однако они не связанны с его основными тенденциями и поэтому могут не приниматься в расчет при его систематических исследованиях. Напротив, ключевую роль здесь должны играть именно стабильные источники, отражающие (и в какой-то мере порождающие) реальные закономерности сетевой динамики.

На практике среди множества проблем подбора и анализа источников контента большое значение имеет учет параметров их стабильности, в частности, тематической. При этом тематическая стабильность и стабильность потока информации от источников зачастую играют решающую роль при проведении аналитических исследований. Например, такие важные свойства информационных источников, как их тематическая корреляция [21] и полнота, 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. Предполагается, что конкретный набор рубрик мало влияет на предлагаемый ниже метод расчета стабильности источников (под тематической рубрикой в данном случае понимается тематика, семантика которой, в частности, находит свое отражение в виде запроса на информационно-поисковом языке). Предполагается, что документу присваивается та или иная рубрика, если он соответствует определенному запросу. Перечень рубрик и соответствующих им запросов был выбран авторами на основании опыта работы с политематическими новостными ресурсами сети Интернет. Эти рубрики и запросы установлены и апробированы в течение

длительного времени в системе контент-мониторинга InfoStream. В настоящее время система включает 35 основных тематических рубрик. При этом именно эта система, охватывающая более 30000 новостных сообщений в сутки, была выбрана в качестве экспериментальной платформы.

При исследовании тематической направленности некоторых источников информации были обнаружены документы, отклоняющиеся от основной направленности этих источников. Такие документы, если их количество относительно невелико, не должны влиять на рассчитываемый ниже уровень стабильности источников. Конечно, автоматическая рубрикация во многом зависит от качества запросов, однако некоторыми погрешностями в рубрикации при статистическом исследовании можно пренебречь.

Для подхода к изучению стабильности источников важно знать параметры их распределения по тематическим рубрикам, т.е. количество рубрик, соответствующих документам, входящим в эти источники. Результаты такого исследования, охватывающего 920 репрезентативных русскоязычных источников (опубликовавших за месяц более 100 сообщений), приведены на рис. 26. Об относительно невысокой тематической стабильности источников, порождающих общий информационный поток системы, свидетельствует тот факт, что около половины репрезентативных источников соответствуют более 20 рубрикам.

Для вычисления уровня разброса (нестабильности) источника информации использовалась формула, основанная на линейной метрике:

$$R = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{M*\max(r_i)}\sum_{j=1}^{M}\left|r_{ij} - \frac{1}{M}\sum_{k=1}^{M} r_{ik}\right| \quad , \qquad (3.8)$$

где N – количество рубрик, M – количество дней, $\max(r_i)$ – максимальное суточное количество вхождений рубрики i в документы источника за все время, r_{ij} –количество вхождения рубрики i за день j.

Из приведенной формулы следует, что значение R, на самом деле, учитывает не только тематический разброс, но и разброс по количеству вхождений рубрики, т.е. фактическое количество документов от источника, относящихся к данной рубрике. Даже если источник соответствует одной

рубрике, но его наполнение не является стабильным, значение может существенно отличаться от нулевого.

Ранговая диаграмма распределения уровня разброса для источников – Web-сайтов, ежедневно публикующих новостные сообщения в ноябре 2005 года, по уровням тематической стабильности приведена на рис. 27.

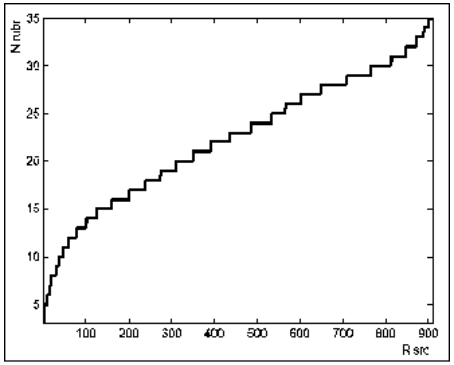


Рис. 26. Ранговая диаграмма "Источники – количество рубрик"

Определение стабильности документов выполнялось по такому алгоритму:

1. Проводился поиск документов в базе данных за определенный период.
2. Формировалась таблица, которая включала код источника информации, коды соответствующих ему тематических рубрик и их количество в разрезе дат.
3. Для каждого источника по приведенной выше формуле определялся уровень разброса R.
4. Информационные источники ранжировались по рассчитанным параметрам, и строилась соответствующая диаграмма.

Как оказалось, источники, содержащие до 5-6 рубрик, обладают исключительной стабильностью, что, в общем, достаточно очевидно. Не совсем очевидным оказался факт резкого всплеска разброса для источников, включающих документы с 25 и более рубриками.

В приведенной для вычисления уровня разброса формуле используется линейная метрика, которая позволяет достаточно точно дифференцировать источники при минимальных затратах на вычисления. Необходимо отметить, что для анализа временных рядов сегодня также широко используются вычисления, базирующиеся на евклидовой метрике. В частности, популярный в экономике подход R/S-анализа позволяет исследовать «изрезанность» кривой, образуемой временным рядом на основе отношения разброса значений к среднеквадратичному отклонению [35]. Очевидно, что «изрезанность» кривой близка по смыслу фрактальной размерности. Заметим, что в то время как фрактальная природа научных коммуникаций исследовалась в работах Ван Раана [64] и С.А. Иванова [12], новостные информационные потоки в сети Интернет до настоящего времени не рассматривалось в научных публикациях с этой точки зрения.

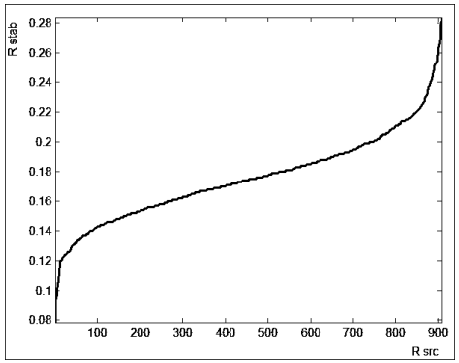


Рис. 27. Диаграмма "Ранг источника - коэффициент разброса"

Аналог формулы (3.8) – мера «изрезанности» распределения временного ряда при фрактальном походе R_f выглядит следующим образом:

$$R_f = \frac{1}{N}\sum_{i=1}^{N}\frac{S_i}{R_i} \quad , \qquad (3.9)$$

где S_i – среднеквадратичное отклонение по рубрике i:

$$S_i = \sqrt{\frac{1}{M}\sum_{j=1}^{M}(r_{ij} - \frac{1}{M}\sum_{k=1}^{M} r_{ik})^2} \quad , \qquad (3.10)$$

R_i – размах значений по рубрике i:

$$R_i = \max_{1<k<M} X_{ik} - \min_{1<k<M} X_{ik} \, , \qquad (3.11)$$

X_{ik} – накопленное к моменту k отклонение по рубрике i:

$$X_{ik} = \sum_{j=1}^{k}(r_{ij} - \frac{1}{M}\sum_{l=1}^{M} r_{il}) \quad . \qquad (3.12)$$

В соответствии с формулой (3.9) коэффициент «изрезанности» временного ряда количества публикаций по N темам одного источника выглядит следующим образом:

$$R_f = \frac{1}{N}\sum_{i=1}^{N}(\frac{2}{M})^{H_i} = \frac{1}{N}\sum_{i=1}^{N}(\frac{2}{M})^{1-\rho_i} \quad . \qquad (3.13)$$

На рис. 28. представлена кривая значений коэффициентов «изрезанности» для источников (было измерено поведение 1700 источников за 2005 год), ранжированных по этим значениям. Как видим, характер кривой вполне соответствует характеру кривой разброса источников (рис. 27). Более того, названные коэффициенты очень близки по своей природе.

Коэффициент Херста для источника i можно вычислить по формуле:

$H_i = log(R_i/S_i)/log(M/2)$. (3.14)

На рис. 29 приведена совместная диаграмма коэффициентов «изрезанности» и коэффициентов Херста.

Из фрактального анализа известно, что коэффициент Херста, равный ½, соответствует броуновскому движению, т.е. случайному поведению временного ряда по тематикам, чем значние коэффициента Херста ближе к единице, тем процесс тематических публикаций более детерминирован, т.е. персистентен. Это означает, что если количество публикаций растет, то можно предположить, что оно будет возрастать и в дальнейшем, а если уменьшается, то и в дальнейшем будет наблюдаться спад. Значение коэффициента Херста меньшее ½

свидетельствует об антиперсистентности или эргодичности процесса. Данное свойство означает, что если количество публикаций растет, то можно предположить, что оно в дальнейшем пойдет на спад, а если уменьшается, то и в дальнейшем будет наблюдаться рост количества публикаций. Конечно, такие источники можно оценивать как крайне тематически нестабильные.

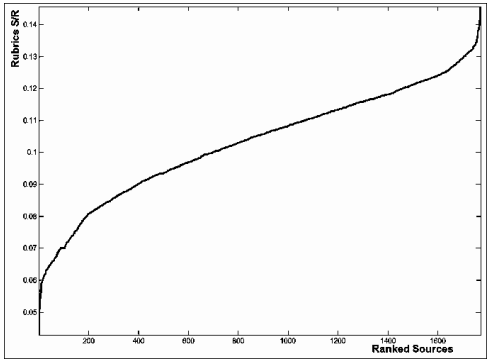


Рис. 28. Диаграмма "Ранг источника - коэффициент «изрезанности»"

При этом рассматривались «усредненные» по количеству тематических рубрик значения коэффициента Херста:

$$H = \frac{1}{N}\sum_{i=1}^{N} H_i \quad , \qquad (3.15)$$

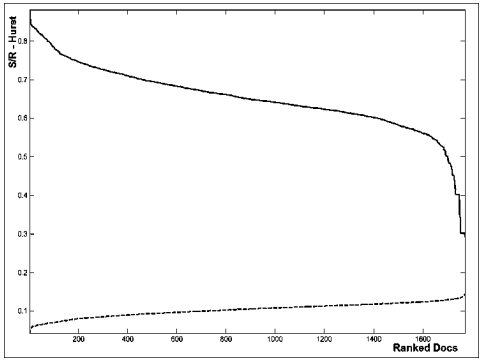


Рис. 29. Диаграмма "Ранг источника – коэффициент тематической «изрезанности» и коэффициент Херста"

Как оказалось, имеет большой практический смысл рассматривать стабильность процесса поступления информации от отдельных источников даже вне тематического разреза. Параметр «изрезанности» кривой количества поступлений от источников вне тематического разреза, конечно, более простой, однако он также может быть определяющим при выборе списка источников абонентами системы контент-мониторинга.

В этом случае:

$$R_f = \frac{S}{R} = \left(\frac{2}{M}\right)^H \quad , \qquad (3.16)$$

где S - среднеквадратичное отклонение, R – размах, H – коэффициент Херста.

На рис. 30. представлена ранговая диаграмма внетематического распределения источников по коэффициентам «изрезанности». Как видно, форма графиков отличается от представленных на рис. 29 практически отсутствием одной из точек перегиба, а значения показателя Херста в среднем существенно выше, что свидетельствует об очевидном факте: электронные издания более склонны изменять тематику своих публикаций, чем их объемы, выраженные количеством публикаций.

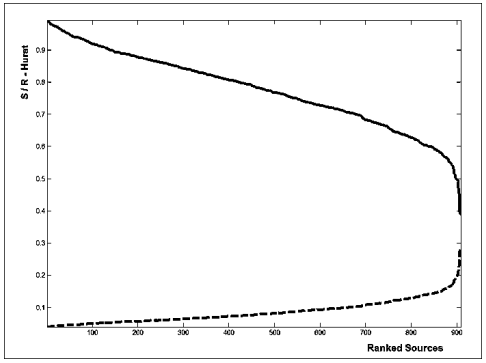


Рис. 30. Диаграмма "Ранг источника – коэффициент внетематической «изрезанности» и коэффициент «Херста»"

Результаты исследований стабильности источников 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 информационных источников» в интерфейсе системы контент-мониторинга информационных ресурсов.

Сегодня становится ясно, что разработка качественно новых средств работы с сетевыми ресурсами переходит в разряд приоритетных задач. Без приемлемых средств контроля за сетевыми информационными процессами невозможно обеспечить репрезентативность выборок. В любом случае, успешное продвижение в изучении современного информационного пространства невозможно без хотя бы общих представлений о структуре и свойствах динамики сетевых информационных процессов, что в свою очередь требует выявления их устойчивых закономерностей.

4. Web второго поколения

За 15 лет своего существования Web-пространство превратилось в крупнейший в мире распределенный информационный ресурс благодаря нескольким заложенным в его основу принципам. К этим принципам относится реализация гипертекста, позволяющего интегрировать неоднородные информационные ресурсы, использование простого, доступного пониманию пользователей языка разметки HTML (что обусловило легкость публикации документов в Сети), естественную, адаптированную к человеческой логике систему навигации в гипертекстовой среде.

Вместе с тем возможности представления и доступа к информации в Web ограничивались статичностью языка HTML, что обуславливало только навигационный доступ к ресурсам, практическое отсутствие поддержки метаинформации, несовершенство идентификации информационных ресурсов, и, самое главное, тот факт, что разметка HTML относилась только к внешнему представлению документов, не касаясь их семантики.

По мере развития Web первого поколения его возможности расширялись, эволюционно были добавлены динамические компоненты, возможность управлять стилевыми решениями, были разработаны и некоторые принципы представления контента, зафиксированные как рекомендации. В процессе этой эволюции появились Java-аплеты и Java-скрипты, многочисленные META-теги, язык каскадных таблиц стилей CSS и т.д.

Вместе с тем традиционному Web все же присущи такие недостатки, как высокий уровень информационного шума, невозможность гарантирования целостности документов, отсутствие возможности смыслового поиска, ограниченность доступа к «скрытому» Web.

Над решением названных проблемам работают многочисленные коллективы ученых и специалистов во всем мире, в частности, консорциум W3C, где под руководством основателя WWW Тима Бернерса-Ли реализуется концепция Семантического Web [42]. Наряду с этой концепцией, революционный

прорыв обещает дать более общий подход, а именно Web-2.0, или как его называют, “Web второго поколения” [63], который включает в себя реализацию концепции семантического Web, многоуровневую поддержку метаданных, новые подходы к дизайну и соответствующему инструментарию, технологию глубинного анализа текстов (Text Mining), а также идеологию Web-сервисов, базируясь на информационных ресурсах, накопленных в Web первого поколения. Таким образом, Web-2.0 предусматривает пересмотр всего комплекса стандартов и архитектурных принципов WWW. Сегодня очевидно, что центральное звено инструментария представления и обмена данными будет играть Расширяемый Язык Разметки (XML) [16], лежащий в основе Семантического Web. Предполагается также использование нового принципа идентификации информационных ресурсов, формирование новой архитектуры Web-пространства на основе многоуровневого представления информационных ресурсов и стандартизованных Web-сервисов. Предполагается, что Web-2.0 в начале будет базироваться на ресурсах (базах данных, сайтах, Интернет-сообществах) популярных Интернет-компаний, таких как Google, Amazon, eBay и др.

Семантический Web

Одна из основных частей Web-2.0, которую ее создатели считают абсолютно самодостаточной, является Семантический Web (Semantic Web). Концепцию Семантического Web выдвинул Тим Бернерс-Ли, один из основоположников World-Wide Web и председатель WWW-консорциума (W3C) на международной конференции XML-2000, прошедшей в 2000 году в Вашингтоне.

Основная идея этого подхода заключается в организации такого представления данных в Сети, чтобы допускалась не только их визуализация, но и их эффективная автоматическая обработка программами разных производителей. Путем таких радикальных изменений концепции традиционного Web предполагается превращение его в систему семантического уровня. Семантический Web должен обеспечить «понимание» информации

компьютерами, выделение ими наиболее подходящих по тем или иным критериям данных, и уже после этого - предоставление информации пользователям [19].

Семантический Web можно представить как симбиоз двух направлений, первое из которых охватывает языки представления данных. На сегодняшний день основными такими языками являются Расширяемый Язык Разметки XML (eXtensible Markup Language) и Средства Описания Ресурсов RDF (Resource Description Framework). Существует также ряд других форматов, однако XML и RDF предоставляют больше возможностей, потому они обладают статусом рекомендаций W3C.

Второе, концептуальное направление несет в себе теоретическое представление о моделях предметных областей, которые в терминологии Семантического Web называются онтологиями. 10 февраля 2004 года консорциумом W3C была утверждена и опубликована спецификация языка сетевых онтологий OWL (Web Ontology Language).

В результате две ветви Семантического Web используют три ключевых языка (соответственно, технологии) [36]:

- спецификация XML, позволяющая определить синтаксис и структуру документов;
- механизм описания ресурсов RDF, обеспечивающий модель кодирования для значений, определенных в онтологии.
- язык онтологий OWL, позволяющий определять понятия и отношения между ними.

Семантический Web использует также и другие языки, технологии и концепции, в частности, универсальные идентификаторы ресурсов, цифровые подписи, системы логического вывода и т. д.

Практическая реализация Семантического Web критически зависит от существования Web-страниц, содержащих метаданные, формирование которых не входит в стандартный процесс Web-разработки. Вряд ли удастся заставить авторов Web-страниц вручную индексировать свои ресурсы с помощью терминологических словарей, онтологий Семантического Web. Очевидно, что

интегрировать существующие ресурсы WWW в Web-2.0 (что предусмотрено базовой концепцией) можно только автоматически. Данная задача является очень сложной, требует подходов технологии глубинного анализа текстов (Text Mining), которая, в свою очередь, сегодня бурно развивается.

Метаданные и онтологии

Современная постановка задачи поиска информации сводится к находжению тех данных, которые действительно необходимы, и преобразованию их в информацию, требуемую конкретному потребителю. В случае успешного ее решения в сфере информационных технологий, несомненно, откроются новые перспективы.

Под таким углом зрения главной задачей становится разработка средств, позволяющих ориентироваться в сложных комплексах разнообразных наборов данных с целью извлечения из них нужной информации. Сложность, однако, заключается в том, что, в отличие от методик работы собственно с данными, которые сравнительно легко формализуются, здесь так или иначе придется осуществить переход от уровня формальных систем к уровню систем содержательных.

Одним (как представляется сегодня, наиболее эффективным) путем реализации подобной программы является использование метаданных.

Данные, вообще говоря, можно определить как совокупность формальных элементов, описывающую свойства (состояние) некоторого объекта. Когда мы говорим о данных, то обычно предполагаем, что этот объект относится к интересующей нас предметной области. Однако ничто не мешает в качестве объекта рассматривать сами данные. Они, разумеется, в свою очередь также могут быть описаны некоторыми данными, которые в этом случае и называют метаданными (исходные данные, соответственно, называют объектными). Таким образом, не будет ошибкой сказать, что метаданные – это данные о данных.

Ясно, что разделение на объектные и метаданные всегда относительно и носит условный характер, приобретая конкретный смысл лишь в определенной

ситуации при определенном ее рассмотрении. Если объектными данными будем считать хранящиеся в библиотеке книги, то соответствующими метаданными будут их карточки в каталоге. Если же в качестве объектных данных мы выберем карточки, то метаданными могут считаться коды ящиков каталога, где они находятся.

Физически и структурно метаданные могут входить в состав соответствующих документов, или же располагаться отдельно, образуя собственные БД, иногда достаточно сложные и разветвленные.

Главное концептуальное свойство метаданных, определяющее их перспективность в плане интересующих нас вопросов, состоит в том, что они могут содержать в себе элементы качественно иной природы, чем элементы объектных данных. Например, указание в библиотечной карточке книги количества ее страниц может облегчить нам работу, но ничего не прибавляет по существу, т. к. имея саму книгу, мы без труда определим это сами. Но вот то, что книга представляет собой перевод с японского, мы вряд ли догадаемся, сколько бы ее ни разглядывали. Вообще, метаданные могут содержать (и, как привило, содержат) характеристики, которые невозможно получить, манипулируя объектными данными.

Для того, чтобы использовать метаданные тем или иным способом, их вначале необходимо создать. В рамках информационных технологий существует два способа, которые, впрочем, могут сочетаться. Метаданные могут создаваться генераторами параллельно объектным данным или формироваться автоматизированно посредством обработки уже готовых объектных данных. На первый взгляд, второй способ выглядит предпочтительнее, однако на наш взгляд, он обладает существенными недостатками, обусловленными в первую очередь тем, что любой автоматизированный процесс неизбежно предполагает достаточно высокий уровень формализации, что снижает эффективность обсуждаемой методики. В этом случае построенные наборы метаданных будут мало отличаться от самих объектных данных. Это, разумеется, тоже может оказаться полезным, но

существенно сужает рамки применимости концептуальных возможностей подхода.

Поэтому мы отдаем предпочтение первому способу. Проще говоря, идея состоит в том, что генерируемые наборы объектных данных сразу же снабжаются сопряженными с ними наборами метаданных, предназначенных для автоматической обработки. Особенно эфективными могут быть наборы метаданных, формируемые опытными экспертами в данной предметной области, способными находить емкие и лаконичные характеристики наиболее сущностных аспектов той или иной информационной единицы.

Конечно, практическое применение подобных подходов требует значительных организационных затрат, по крайней мере на первых этапах внедрения в реальные технологические процессы. Однако конечный результат, без сомнения, окупит все технические сложности.

Приведем основные направления, в которых, на наш взгляд, описанная методика может дать реальную отдачу в ближайшем будущем.

Прежде всего, это информационный поиск. Здесь возможны две основные модификации: собственно поиск по базам метаданных и использование метаданных для конструирования различных фильтров, сужающих релевантную выборку.

В первом случае вместо составления запроса из поисковых терминов, которые должны присутствовать в тексте релевантного документа, следует указать интересующие вас характеристики, содержащиеся в наборах метаданных. Например, английские детективы первой половины XX-го века. Конечно же, непосредственно в текстах самих детективов не указано, что они детективы, а уж принадлежность к той или иной половине XX-го века и подавно. В то же время правильно организованные метаданные такие сведения содержать могут («Жанр», «Страна издания», «Дата издания» и т. д.).

Во втором случае с помощью метаданных пользователь может задать ряд условий, которым должен удовлетворять релевантный документ, отобранный по обычному запросу. Например, вы задаете запрос «топология банахова

пространства», но при этом хотите получить только академические монографии. Очевидно, что никакими изощренными комбинациями поисковых терминов вы не сможете отличить монографию от статьи, а специальную статью, в свою очередь, от популярной. Зато с помощью надлежащим образом построенных наборов метаданных отобрать требуемые документы по данному критерию не составит труда.

Оба способа могут применяться совместно в различных «смешанных» вариантах: использование различных наборов метаданных в различных комбинациях создает широкое пространство для оптимизации поисковых процедур. При этом механизм отбора релевантных документов становится качественно иным, поскольку подключаются возможности использования параметров, вообще не связанных непосредственно с текстами.

Другой сферой применимости метаданных к информационным технологиям является задача кластеризации документальных массивов. Общий принцип остается тем же, только используется он для группировки документов по заданным признакам. Именно эти признаки и задаются с помощью соответствующих наборов метаданных. Частным случаем кластеризации в этом смысле можно также считать задачу избирательного распространения информации. Действительно, соотнесение документа с тем или иным каналом распространения принципиально не отличается от сказанного выше.

В заключение приведем список основных преимуществ технологий, основанных на использовании метаданных.

- Метаданные, дополняющие полнотекстовые БД, позволяют использовать общую идеологию реляционной модели вместе со всеми обширными наработками (сами документы тогда играют роль MEMO-полей).
- Метаданные позволяют формировать такие комплексы характеристик документов, какие в принципе невозможно зафиксировать с помощью запроса, поскольку в них используются понятия, отсутствующие в самих текстах документов.

- Метаданные могут включать признаки, актуальные для потребителя, но при этом вообще не связанные со свойствами самих документов (например, рецензии на статьи, содержащие оценки их профессионального уровня).
- Метаданные позволяют осуществлять кластеризацию выборки документов по признакам, определяющим внешний контекст (например, публикации, вышедшие до роспуска Парламента и после).
- Метаданные позволяют использовать в процессе поиска методы агрегирования.
- При создании наборов метаданных могут использоваться технологические возможности, недоступные генераторам документов в силу их недостаточного профессионального уровня.

С концепцией метаданных неразрывно связаны онтологии - один из важнейших компонент Семантического Web. В философии онтологией называют теорию о природе бытия и видах сущностей. Онтологический уровень формализует накопленные знания, определяя и объединяя терминологию различных предметных областей.

Онтологии получили достаточно широкое распространение в задачах представления знаний и инженерии знаний, семантической интеграции информационных ресурсов, информационного поиска и т.д. В науке об «искусственном интеллекте» онтология – это "спецификация концептуализации предметной области", или упрощенно, документ, формально задающий отношения между терминами. Это своего рода словарь понятий предметной области и совокупность явным образом выраженных предположений относительно смысла этих понятий.

Чаще всего онтология представляется как иерархия понятий, связанных отношениями некоторых определенных видов. Такие определения онтологий используются в различных классификациях. Развитые онтологии формализуются средствами языков логики и допускают возможности логического вывода.

В простейшем случае онтологии можно использовать для повышения точности поиска в Интернет - поисковая система будет выдавать только такие сайты, где упоминается в точности искомое понятие, а не произвольные страницы, в тексте которых встретилось заданное ключевое слово.

Согласно принципам Семантического Web, процесс создания электронных документов включает два этапа: создание документа, содержащего некоторые термины и создание его онтологии. Онтология может описываться различными средствами и сегодня существует несколько языков описания онтологий, однако, точная семантика описываемых терминов и связей в различных языках одна и та же.

Общеизвестно, что в различных предметных областях одни и те же понятия могут быть представлены разными терминами. Механизм онтологий в этих случаях позволяет формировать осмысленные иерархические взаимосвязи между объектами, обобщать и совместно использовать глобальные сведения, т.е. реализовать нечеткий поиск, способный находить даже такие необходимые пользователю ресурсы, в которых не будет ни одного слова из исходного запроса.

Предполагается, что «интеллектуальные» приложения смогут использовать онтологии, чтобы получать в результате поиска информацию со связанной с ней структурой знаний и правилами вывода. Механизмы поиска могут применять онтологии и для выборки страниц с синтаксически различными, но семантически одинаковыми словами. Онтологии также могут использоваться для организации обмена данными и интеграции программ.

Такая интеллектуальная программа, интерпретирующая онтологии, сможет вывести, например, что если Корнелльский Университет находится в г. Итака, который расположен в штате Нью-Йорк, который, в свою очередь, входит в США, то адрес этого университета следует писать в американском формате.

Разработка языка описания структурированных онтологий OWL стало в последнее время одним из наиболее важных звеньев работ по Семантическому Web, проводимых консорциумом W3C. В конце 2001 года для этой цели в составе W3C была учреждена специальная рабочая группа – Web Ontology Working

Group. 10 февраля 2004 года WWW-Консорциум присвоил языку OWL статус рекомендованной к реализации технологии.

В рамках OWL онтология – это совокупность утверждений, задающих отношения между понятиями и определяющих логические правила для рассуждений о них. Онтология может включать описания классов, свойств и их примеры. Компьютеры могут «понимать» смысл семантических данных на Web-страницах, следуя по гиперссылкам, ведущим на онтологические ресурсы. Онтология может включать описания классов, свойств и их примеры (индивиды).

OWL может использоваться, чтобы явно представлять значения терминов и отношения между этими терминами в словарях. OWL имеет больше средств для выражения значения и семантики, чем XML, RDF, и RDF-S, и, таким образом, OWL идет дальше этих языков в способности представить поддающийся машинной обработке контент Сети. С другой стороны, OWL обеспечивает более полную компьютерную обработку Web-контента, чем та, которую поддерживают XML, RDF, и RDF Schema (RDF-S), предоставляя наряду с формальной семантикой дополнительный терминологический словарь.

Формальная семантика OWL описывает, как получить логические выводы на основе онтологий, т. е. получить факты, которые не представлены буквально, а следуют из семантики онтологии. Эти выводы могут базироваться на анализе одного документа или множества документов, распределенных в Сети.

Последнее обеспечивается возможностью онтологий быть связанными, включая прямой импорт информации из других онтологий. Чтобы написать онтологию, которая может однозначно интерпретироваться и использоваться программными агентами, задействуются синтаксис и формальная семантика OWL.

В языке OWL, прежде чем использовать какое-либо множество терминов, необходимо точно указать словари (онтологий), которые записываются в формате URI. Благодаря этому отпадает необходимость создания дополнительной базы данных, содержащей имена всех онтологий — роль базы данных уникальных идентификаторов возлагается на распределенную базу данных доменных имен.

OWL предполагает открытость, т.е. описания ресурсов не ограничены единственным файлом или темой. В то время как некоторый класс первоначально может быть определен в онтологии, он может быть расширен в других онтологиях. Следствия из этих дополнительных суждений о заданном в начале классом являются монотонными. Новая информация не может опровергать предыдущую информацию. Новая информация может быть противоречащей, поэтому в OWL факты и следствия могут только добавляться, и не могут удаляться.

На практике создание онтологий начинается с иерархии классов понятий, составляющих предметную область. Наиболее фундаментальные понятия в какой-то области должны соответствовать классам, которые находятся в корне различных таксономических деревьев. Каждый индивид в мире OWL является членом класса owl:Thing. Таким образом, каждый определенный пользователем класс автоматически является подклассом owl:Thing. Специфичные для данной области корневые классы определяются простым объявлением именованного класса. OWL также определяет пустой класс, owl:Nothing. Фундаментальным таксономическим конструктором для классов является rdfs:subClassOf. Он связывает частный класс с более общим классом. Если X - подкласс Y, то каждый представитель X - также представитель Y. Отношение rdfs:subClassOf является транзитивным. Если X - подкласс Y, и Y - подкласс Z, то X - подкласс Z.

Свойство в OWL - это бинарное отношение. Различают два типа свойств:

- свойства-значения, отношения между представителями классов и RDF-литералами или типами данных, определяемых XML Schema;
- свойства-объекты, отношения между представителями двух классов.

В OWL существует множество способов ограничения отношений.

По словам Тима Бернерса-Ли, приведенным в пресс-релизе W3C, RDF и OWL — это серьезный шаг и весьма мощная база для приложений Семантического Web. Утверждение этих стандартов как рекомендованных W3C-консорциумом подоспело вовремя. Сегодня открывается новая фаза Internet как

информационного пространства. Эта фаза началась с того момента, когда проект Семантический Web начал свою работу.

Поисковые системы

Поскольку количество Web-сайтов продолжает стремительно увеличиваться, пользователи Web-2.0 нуждаются в более эффективных поисковых системах. Поисковые машины следующих поколений должны будут лучше классифицировать информацию и нагляднее представлять ее. В будущем поиск не будет ограничиваться лишь обработкой введенных ключевых слов. Например, во внимание будет приниматься местоположение пользователя. Системы станут отслеживать интересы пользователей, делая поиск более целенаправленным. Новое программное обеспечение будет работать с мультимедийной информацией так же легко, как с текстом. Новые поисковые машины будут "видеть" опубликованные в Сети текстовые, аудио- и видеоматериалы, которые в настоящее время недоступны.

В последнее время получили распространение адаптивные интерфейсы уточнения запросов [4], чаще всего реализуемые путем кластеризации результатов первичного поиска. Появилось такое понятие, как метод "папок поиска" (Custom Search Folders), который не связывается с определенным алгоритмом кластеризации, а представляет собой множество подходов, общее у которых - попытка сгруппировать результаты поиска и представить кластеры в удобном для пользователей виде.

К подобным механизмам можно отнести, например, австралийский поисковый сервер Mooter (http://www.mooter.com/), на котором применяется визуальный подход к предоставлению результатов поиска по обрабатываемым запросам путем группировки результатов первичного поиска по категориям. Другой поисковый сервер iBoogie (http://www.iboogie.com/) также группирует результаты поиска, но отображает их в виде, близком к экрану проводника Windows. Слова и словосочетания в информационных портретах, применяемых, например, в системе Галактика Зум, также позволяют адаптивно уточнять

первичные запросы. Недавно разработчики Google представили свои наработки и планы по кластеризации найденных документов. Демо-версия этой системы позволяет выделять из документов названия компаний, которые являются основными критериями кластеризации.

Одним из наиболее интересных решений следует считать метод так называемых информационных портретов, использующихся для уточнения запросов. При этом уточнение осуществляется за счет добавления не произвольных терминов, придумываемых пользователем, а определенного их набора, формируемого машиной в процессе статистической обработки доступного массива данных. Иными словами, пользователю предлагается то, что реально существует (он может обнаружить в списке термин, вполне отвечающий его потребностям, но который сам он не смог бы придумать). Поэтому, возможно, правильнее было бы говорить не об уточнении поиска, а о его сужении.

Воплощением идеи коллективной работы в Интернет, входящей в концепцию информационно-поисковых систем нового поколения, сегодня стала система с "хвостовыми данными" Snap, обеспечивающая не только поиск Web-страниц по ключевым словам, но и предоставляющая дополнительную информацию, близкую интересам пользователей. Например, к результату поиска по изготовителям цифровых камер добавляется сравнительная таблица моделей, которые ранее были затребованы другими пользователями системы. Данная поисковая система является предвестником такого этапа развития WWW, на котором в ней будут активно использоваться результаты работы всего сообщества пользователей.

В Информационном центре «ЭЛВИСТИ» была разработана система InfoStream [5], которая применяется для решения задач автоматизированного сбора новостной информации с Web-сайтов, ее обработки и обеспечения доступа к ней в поисковых режимах. Эта система охватывает свыше 2500 Web-источников – более 40000 уникальных новостных сообщений в сутки, при этом в ретроспективных базах данных хранится свыше 30 млн. сообщений. На рис. 31 представлен пример сужающего списка терминов, используемого нами при

обработке потоков новостной информации в рамках оригинальной технологии InfoStream.

Для эффективной работы с такими объемами информации простейшего информационного портрета оказалось мало – понадобился «информационный альбом» - многоаспектная подборка параметров выборки по первоначально составленному запросу. И такая возможность была реализована. При этом в отличие от большинства подобных систем, в InfoStream уточняющие параметры поиска задаются не заполнением сложной формы расширенного поиска, а указываются путем выбора из информационного альбома, получаемого в результате поиска по первичному запросу.

Конечно, новые возможности потребовали существенного пересмотра концепции индексирования документов, выбора из текстов документов и нормализации ключевых слов-дескрипторов, выявления ряда содержательных параметров документов.

Сегодня в системе InfoStream информационный альбом, соответствующий первичному запросу, содержит такие параметры, как ключевые слова, рубрики, языки, страны. В частности, в адаптивном интерфейсе системы существенно облегчен множественный выбор источников информации, соответствующих заданному запросу. Кроме того, пользователю предоставлена возможность задания характеристик размеров искомых документов. Предусмотрен и такой «экзотический» параметр, как уровень насыщенности документов цифровой информацией, что полезно, например, при поиске аналитических документов, ценовых таблиц, рейтингов и т.п.

Экспериментальной реализацией идеи коллективной работы в Интернет, входящей в концепцию Web-2.0, стала поисковая система Snap (http://www.snap.com/), обеспечивающая не только поиск Web-страниц по ключевым словам, но и предоставляющая дополнительную информацию, близкую интересам пользователей.

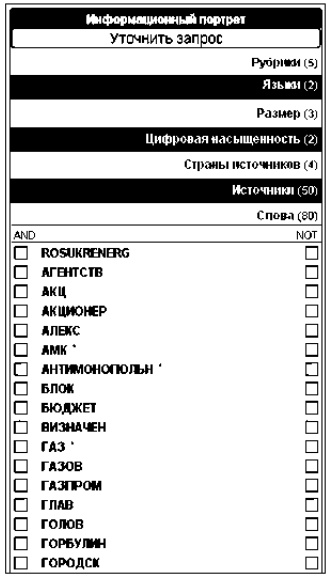


Рис. 31. Информационный альбом системы InfoStream

Например, к результату поиска по изготовителям цифровых камер добавляется сравнительная таблица моделей, которые ранее были затребованы другими пользователями системы. Разработчик системы Билл Гросс считает, что Snap стала первой системой с "хвостовыми данными" (data trail), но вскоре таких систем станет больше. Данная поисковая система является предвестником такого этапа развития WWW, на котором в ней будут активно использоваться результаты работы всего сообщества пользователей.

Новые поисковые системы улучшают качество результатов, все глубже зарываясь в доступные хранилища информации, сортируя ее, и представляя результаты с учетом персональных пользовательских предпочтений. Так, недавно порталы Amazon, Ask Jeeves и Google объявили о внедрении механизма улучшения результатов поиска, основанного на персонализации. Поисковые машины www.A9.com (проект Amazon) и www.MyJeeves.ask.com (проект Ask Jeeves) не только отслеживают запросы и найденные Web-страницы, но также позволяют сохранять их в виде закладок. Пользователь MyJeeves может

многократно просматривать накопленные результаты, которые представляют собой персонально организованную область WWW. Подобные функции поддерживает и портал www.A9.com, на котором предлагается набор страниц, сформированный при анализе личной поисковой истории. Истории поисковых запросов на сайтах A9 и MyJeeves хранятся на серверах поисковых систем. В системе Google пользователь может выбрать из иерархического списка наиболее важные для него темы и указать степень своего интереса к той или иной области знаний. Все эти данные учитываются системой при оценке результатов поиска.

«Скрытый» Web

Web-2.0 предполагает открыть доступ к “скрытому” Web. Большая часть содержания сайтов WWW первого поколения остается недоступной для поисковых машин, потому что многие Web-серверы хранят и перерабатывают информацию не в том виде, в каком она представляется посетителю. При этом многие Web-страницы генерируются только тогда, когда пользователи обращаются к ним. Традиционные сетевые агенты не умеют работать с подобными ресурсами и не в состоянии определить их содержание. «Скрытый» Web охватывает в первую очередь содержимое онлайновых баз данных [15]. Скрытой является и быстро обновляемая информация - новости, конференции, онлайновые журналы.

В 2000 году американская компания BrightPlanet (www.brightplanet.com) опубликовала сенсационный доклад [68], в котором утверждается, что в WWW в сотни раз больше страниц, чем их удалось проиндексировать самыми популярными поисковыми системами.

На сегодня разработан целый класс программ, получивших название упаковщиков (wrappers). В некоторых программах, чтобы получить доступ к скрытому содержанию Web-страниц, используется привычный синтаксис поисковых запросов и стандартный формат онлайн-ресурсов. В других системах реализуются преимущества программируемого интерфейса, который позволяет использовать стандартный набор команд и операций.

Для поиска в "скрытой" Сети, а именно в том ее сегменте, который составляют базы данных, сегодня уже существуют некоторые специализированные ресурсы. Среди них, например, системы BigHub (www.bighub.com) и InvisibleWeb (www.invisible-web.net) компании IntelliSeek. Сайт Invisible Web включает в себя каталог баз данных, большинство из которых не проиндексированы известными поисковыми машинами. При введении запроса этот сайт выдает ссылки на ресурсы, с помощью которых поиск необходимой информации станет наиболее оптимальным. На этом сайте Криса Шермана (Chris Sherman) и Гари Прайса (Gary Price) собраны коллекции ссылок на различные базы данных, среди которых содержится немало уникальных ресурсов, например, сборник спичей политиков и бизнесменов. Программный пакет BullsEye компании IntelliSeek осуществляет поиск более чем в 800 сетевых ресурсах.

В 2005 году компания Yahoo также запустила тестовую версию поискового сервиса, ориентированного на работу с базами данных сайтов. Он обеспечивает проведение поиска не только в общедоступных сайтах, но и на ресурсах, предоставляющих платную информацию, - таких, как онлайновая версия Wall Street Journal, взимающий с посетителей определенную плату. Новый сервис получил название DeepWeb и доступен пока что только для жителей США и Великобритании.

Но все же лидером среди навигаторов в «скрытом» Web является сайт CompletePlanet (www.completeplanet.com) компании BrightPlanet. Этот сайт - крупнейший каталог, насчитывающий 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 «скрытого» Web. Метопоисковый пакет DeepQueryManager (DQM) этой же компании обеспечивает поиск по 55 тысячам "скрытым" Web-ресурсам.

Text Mining

Поисковые технологии Web-2.0 должны стать более эффективными за счет мощных технологий, объединяющих поиск и глубинный анализ текстов (Text Mining), нахождение аномалий и трендов в текстах. Одновременно эти технологии будут неявными, благодаря операциям интеллектуального поиска, встроенным в интерфейсы "по умолчанию". В итоге поиск информации в Web-2.0 станет неразрывно связанным с ее осмыслением.

Существует четыре основных вида приложений технологий Text Mining, которые должны найти свое воплощение в Web второго поколения [18]:

- Классификация текста, в которой используются статистические корреляции для построения правил размещения документов в предопределенные категории.
- Кластеризация, базирующаяся на признаках документов, использующая лингвистические и математические методы без использования предопределенных категорий. Результат - таксономия или визуальная карта, которая обеспечивает эффективный охват больших объемов данных.
- Семантические сети или анализ связей, которые определяют появление дескрипторов (ключевых фраз) в документе для обеспечения навигации.
- Извлечение фактов предназначено для получения некоторых фактов из текста с целью улучшения классификации, поиска и кластеризации.

Недавно разработчики Google представили свои наработки и планы по кластеризации найденных документов в рамках технологии Text Mining. Демо-версия этой системы позволяет выделять из документов названия компаний, которые являются основными критериями кластеризации.

Можно назвать еще несколько задач технологии Text Mining, например, прогнозирование, которое состоит в том, чтобы предсказать по значениям одних признаков объекта значения остальных. Еще одна задача — нахождение аномалий, т.е. поиск объектов, которые своими характеристиками сильно выделяются из общей массы. Для этого сначала выясняются средние параметры

объектов, а потом исследуются те объекты, параметры которых наиболее сильно отличаются от средних значений. Подобный анализ часто проводится после классификации, для того чтобы выяснить, насколько последняя была точна.

Несколько отдельно от задачи кластеризации стоит задача поиска связанных признаков (полей, понятий) отдельных документов. От предсказания эта задача отличается тем, что заранее не известно, по каким именно признакам реализуется взаимосвязь; цель именно в том и состоит, чтобы найти связи признаков. Эта задача сходна с кластеризацией, но не по множеству документов, а по множеству присущих им признаков.

И наконец, для обработки и интерпретации результатов Text Mining большое значение имеет визуализация. Визуализация в Web-2.0 на основе систем Text Mining предполагается как средство представления контента всего массива документов, а также для реализации навигационного механизма, который может применяться при исследовании документов и их классов.

Web-сервисы

Одним из ключевых элементов Web-2.0 являются Web-сервисы - автономные, модульные приложения, предназначенные для реализации информационных процессов в Сети (в частности, бизнес-процессов [38]). Web-сервисы опираются на ряд отраслевых стандартов: WSDL (для описания), UDDI (для информирования и публикации) и SOAP (для обмена сообщениями).

В августе 2002 года, осознав сложность обращения к Web-сервисам в синхронной и асинхронной средах, корпорации BEA, IBM, Microsoft, SAP и Siebel в результате совместных усилий разработали язык реализации бизнес-процессов для Web-сервисов (Business Process Execution Language for Web Services, BPEL4WS или просто BPEL). Язык BPEL позволяет описывать бизнес-процессы и то, как они связаны с Web-сервисами, а также, как бизнес-процессы используют Web-сервисы для достижения поставленных задач. BPEL можно рассматривать как декларативно-процедурный язык программирования. BPEL фактически

представляет собой диалект языка XML. Как и в любом языке программирования, в BPEL определены зарезервированные слова (теги XML):

- Вызов операции с помощью Web-сервиса (<invoke>).
- Ожидание внешнего сообщения (<receive>).
- Генерация ответа для входных/выходных данных (<reply>).
- Ожидание в течение некоторого времени (<wait>).
- Копирование данных между позициями (<assign>).
- Индикация ошибки или сбойной ситуации (<throw>).
- Остановка реализации всего сервиса (<terminate>).
- Отсутствие действий (<empty>).
- Определение последовательности выполнения действий (<sequence>).
- Ветвление с помощью оператора выбора (<switch>).
- Определение цикла (<while>).
- Выполнение одного из нескольких альтернативных маршрутов (<pick>).
- Индикация того, что шаг должен быть выполнен параллельно (<flow>).
- Индикация обработки ошибочной логики с помощью <throw> и <catch>.

В настоящее время уже существует множество Web-сервисов, однако у других программам нет возможности разыскать в Сети Web-сервис, выполняющий ту или иную функцию. Необходимый для повышения эффективности работы Web второго поколения процесс, называемый обнаружением сервисов, станет возможным лишь после того, как широко распространится приведенный выше унифицированный язык, позволяющий описывать сервисы, с тем чтобы агенты могли «понимать», что позволяет делать данный сервис и каким образом им пользоваться. Например, в рамках Семантического Web агенты производителя сервиса и агенты его пользователей могут достичь понимания друг друга путём обмена онтологиями, содержащими необходимые для общения терминологические словари. Более того, агенты смогут даже сами, находя новые онтологии, совершенствовать свои алгоритмы.

Семантика языка описания сервисов (например, BPEL) позволяет агенту описывать, какие именно функции он может выполнять и какие входные данные

ему требуются. Технология обнаружения Web-сервисов сразу же найдет своих пользователей. Например, в сфере малого бизнеса станет гораздо проще налаживать проведение трансакций в области электронной коммерции, имеющих большую степень защиты и автоматизации.

RSS

В Web второго поколения информация как бы «отчуждается» от источника. Соответственно, предусматривается широкое применение формата RSS (Really Simple Syndication, Rich Site Summary, RDF Site Summary), специально предназначенного для легкого и быстрого обмена содержанием Web-сайтов [24]. RSS обеспечивает согласованный способ резюмировать содержимое Web-сайтов, а кроме того, его применение позволяет администраторам сайтов новостей, онлайновых дневников (weblog), форумов и других часто обновляемых Web-ресурсов получать простой унифицированный метод подачи информации о происходящих событиях.

Сегодня RSS принято рассматривать в первую очередь как формат, предназначенный для публикации и обеспечения экспорта новостей на новостных сайтах. После того, как информация преобразована в формат RSS, программа, ориентированная на этот формат, может загружать сведения об обновлениях Web-сайтов, и, в зависимости от результата, выполнять определенные действия, например, автоматически обновлять список актуальных информационных сообщений.

О перспективности RSS уже сегодня свидетельствуют и попытки использования ее в рекламном бизнесе. На конференции Web 2.0, которая проходила в Сан-Франциско, один из руководителей компании Yahoo Дэн Розенвейг (Dan Rosensweig) заявил, что их система контекстной рекламы Overture будет экспортировать ссылки в RSS-каналы.

Дизайн

Web-2.0 рассматривается в первую очередь как эффективная среда для работы с контентом. Естественно, новая концепция выдвигает новые требования к средствам визуализации информации, дизайну. Сегодня создание инструментария дизайнеров для работы с Web второго поколения является передовым фронтом внедрения технологий Web-2.0. Уже сегодня создаются интерфейсы, которые агрегируют информацию из тысяч источников. Так, Amazon.com (http://www.amazon.com/) дает доступ к своей базе данных через открытый API. Каждый желающий может создать персонализированный, более дружественный, по его мнению, интерфейс пользователя, обладающий функциональностью сайта-первоисточника (например, Amazon Light, http://www.kokogiak.com/amazon/).

В качестве основного языка разметки Web-страниц предполагается использовать XML. Прежние языки разметки - HTML и XHTML - решали преимущественно задачи отображения информации, в то время как XML предназначен для ее описания.

Предполагается, что в Web-2.0 будет реализована персонализированная, независимая навигация и управление Web-сайтом. Т.е. пользователь сам сможет контролировать визуальный интерфейс.

Вместе с тем предполагается, что дизайнер ресурсов Web-2.0 будет в большей мере программистом, с помощью инструментальных средств определяющим элементы структуры, навигации и дизайна Web-сайта[2]. Технологичные, интуитивные интерфейсы, - это то, к чему должны стремиться дизайнеры сайтов Web-2.0. В качестве наиболее технологичных сайтов нового поколения можно назвать картографический сервис Google Maps (http://maps.google.com/), фотосервис Flickr (http://www.flickr.com/), а также Интернет-сообщество Del.icio.us (http://del.icio.us/).

Перспективы

Предполагается, что следующая ступень развития WWW будет определяться технологиями работы с огромным объемом информации,

накопившимся в Сети. В частности, предполагается, что Web второго поколения должен характеризоваться переходом от сети документов к сети данных, которые при необходимости агрегируются в семантически связанные документы с помощью Web-сервисов нового поколения. Предполагается существование единого информационного пространства в виде множества единиц данных, которые могут размещаться на многочисленных сайтах. Пользователь будет получать документ путем агрегирования у себя на компьютере информационных единиц, распределенных в Интернете.

Перспективы Web-2.0 будут во многом зависеть от инфраструктуры, в рамках которой будут работать программные продукты со стороны Web-серверов и пользователей. По мнению многих ученых и участников Интернет-рынка, Web второго поколения будет в большей мере, чем сегодня приспособлен для автоматизированной обработки, использования компьютерами. Благодаря этому потребители будут иметь дело с информацией, собранной ведущими информационными компаниями, и создавать новые сервисы.

5. Интеграция информационных потоков

Традиционным подходам к организации поиска сетевой информации присущи такие недостатки, как низкая оперативность, зависимость от выбора источников и ограниченность спектра этих источников, недостаточные поисковые возможности, отсутствие средств уведомления о появлении новых данных, слабая защита компьютерной информации. Оптимальное решение, способное помочь ориентироваться в новостной информации из Интернета, в настоящее время предоставляют информационные службы нового типа — системы интеграции новостей в Web-пространстве.

Технология мониторинга и последующей синдикации интернет-новостей подразумевает такие основные этапы, как "обучение" программ сбора информации структуре выбранных источников, сканирование информации, ее нормирование, приведение к внутрисистемному формату (в последнее время все чаще к XML и его диалекту RSS), классификация, кластеризация, доставка пользователям различными каналами: e-mail, WWW, Wap, SMS, другие приложения. В качестве таких приложений могут выступать, например, ставшие уже традиционными полнотекстовые поисковые системы, а также системы контент-анализа и "глубинного анализа текстов" (Text Mining), используемые для автоматического выявления смысла в текстах.

5.1. Технология интеграции информационных потоков

Решение задач сбора, систематизации и обобщения информации реализуются в виде комплексов контент-мониторинга, которые выполняют основную “черновую” работу по сбору информации из Интернет и обеспечивают создание документальных хранилищ, соответствующих информационным потребностям потребителей. Такие комплексы обеспечивают:

- постоянное пополнение хранилища оперативными сообщениями,

- эффективный одновременный доступ к базам данных со стороны многих пользователей,
- удобные средства поиска необходимой информации.

Рассмотрим основные компоненты систем интеграции новостей на примере комплекса контент-мониторинга InfoStream®. Этот комплекс реализует полномасштабное информационное хранилище, которое учитывает особенности проблематики пользователей, накапливает и надежно хранит информацию для использования в аналитической работе.

Типовой комплекс контент-мониторинга содержит такие автоматизированные подсистемы (центры, рис. 32):

- сбора и обработки информации;
- обеспечения доступа к полнотекстовым базам данных;
- анализа и обобщения информации.

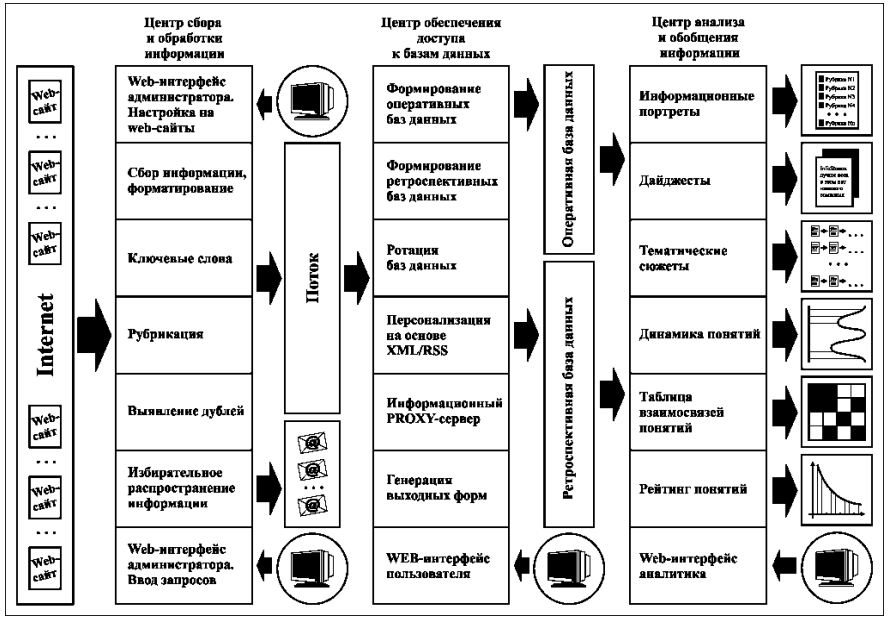


Рис. 32. Основные технологические процессы комплекса контент-мониторинга

В соответствии своему основному назначению, первый центр обеспечивает:

- сбор информации с разнообразных Web-сайтов и ее форматирование;
- выявление ключевых слов, понятий;
- автоматическую рубрикацию;
- выявление содержательного дублирования документов;
- избирательное распространение информации.

Главная задача второго центра - формирование баз данных и обеспечение доступа к ней пользователей, а именно:

- формирование оперативных и ретроспективных баз данных;
- ротация баз данных;
- персонализация работы пользователей, сохранение их персональных запросов и источников, ведение статистики работы;
- обеспечение поиска в базах данных;
- генерация выходных форм;
- информационное взаимодействие с базами данных других подсистем.

Подкомплекс анализа и обобщения информации обеспечивает:

- формирование информационных портретов;
- формирование дайджестов;
- выявление тематических сюжетов;
- построение таблиц взаимосвязи понятий;
- расчет рейтингов понятий.

5.2. Языковые средства интеграции Web-контента

Сегодня доступ к динамической составляющей Интернет – новостным ресурсам – затруднен. Разнообразие информации, в том числе и новостных сообщений, в Сети не может быть полезным на практике при отсутствии эффективного доступа. Так, по оценкам экспертов, около 80% журналистов обращаются к Интернет в поисках новостей, и лишь 20% находят ту информацию, которая им необходима.

Язык HTML, основной формат представления информации в Интернет, описывает лишь внешний вид Web-сайтов, обеспечивая прежде всего

визуализацию данных. Он был разработан исключительно для отображения содержания сайтов, и не всегда удобен для автоматической обработки информации, в том числе и для организации поиска. То есть, вся сеть Интернет ориентирована на показ пользователям отдельных сайтов и плохо приспособлена для автоматизированного сбора информации, ее классификации и аналитической обработки. Сегодня представление информации на разных сайтах настолько отличается по оформлению и расположению, что отбирать ее и обрабатывать можно только вручную.

Так, при необходимости обмена информацией между несколькими Web-сайтами, всегда возникает задача унифицированного представления контента. В противном случае, изменение HTML-оформления одного сайта приведет к необходимости одновременной модификации программного обеспечения на всех сайтах, которые принимают его информацию. Аналогичная ситуация возникает при необходимости импортировать информацию на один сайт с нескольких других. Изменение оформления на каждом из сайтов-источников информации будет всегда приводить к необходимости модификации соответствующего программного кода на целевом сайте.

Как видно, сегодня необходимо использование унифицированного формата данных на сайтах, стандарта, обеспечивающего однотипный обмен данными в Интернет. В качестве такого унифицированного формата все шире используется язык eXtensible Markup Language (XML) и его диалекты.

XML – одна из основных составляющих Семантического Web - представляет собой метаязык, то есть язык, на базе которого можно определять новые языки. При этом он предназначен не только для организации обмена данными в Web, но и для распознавания семантики этих данных. В отличие от HTML, XML обеспечивает представление информации в чистом виде, предполагая ее структурную, а не оформительскую разметку.

Оптимальное решение, способное помочь ориентироваться в новостной информации Интернет, сегодня предоставляют информационные службы нового типа - системы синдикации новостей. Под синдикацией в данном случае

понимается сбор информации в Интернет и последующее распространение ее фрагментов в соответствии с потребностями пользователей. Кроме того, службы синдикации обеспечивают публикацию одних и тех же данных на различных сайтах, в том числе, предназначенных для карманных компьютеров и мобильных телефонов.

Технология синдикации Интернет-новостей включает в себя "обучение" программ сбора структуре выбранных источников (Web-сайтов), непосредственное сканирование информации, ее приведение к общему формату (в последнее время - к XML), классификацию и доставку пользователям различными путями (e-mail, Web, WAP, SMS и т.д.).

Для решения задачи синдикации новостей было создано несколько форматов описания данных на основе XML. Самый распространенный формат получил название RSS, что означает Really Simple Syndication, Rich Site Summary, хотя изначально он назывался RDF Site Summary [25]. Смысл всех этих аббревиатур заключается в простом способе обобщения и распределения информационного наполнения Web-сайтов - синдикации контента.

Изначально RSS создавался компанией Netscape для портала Netcenter как одно из первых XML-приложений, но затем стал использоваться на многих других сайтах. Сегодня практически все ведущие новостные сайты, «живые журналы», работающие в Интернет, используют RSS в качестве инструмента оперативного представления своих обновлений. Например, сегодня экспорт в RSS осуществляют крупнейшие порталы, включая CNN, BBC News, Amazon, CNet News, MSNBC, The Register, Wired и т.д.

RSS действительно обеспечивает согласованный способ резюмировать содержимое Web-сайтов. Кроме того, его применение позволило администраторам новостных сайтов, онлайновых дневников - блогов, форумов и других часто обновляемых Web-ресурсов, представить информацию в унифицированном виде. Сегодня для работы с новостями в формате RSS разрабатываются все новые программы, сайты и поисковые системы, которые все более востребованы.

Итак, RSS - это формат данных и технический стандарт, который обеспечивает интегрированный доступ к новостной информации, представленной на Web-сайтах, специально созданный для обмена их контентом.

Развитие RSS началось с версии 0.90, разработанной компанией Netscape, но ее посчитали очень сложной, и Netscape разработала упрощенную версию - 0.91, которую после бума портальных технологий передала компании UserLand Software. Это самый простой и доступный стандарт, который применяется сегодня в тех ситуациях, когда требуется несложный экспорт заголовков. Одновременно еще одна организация – RSS-DEV Working Group, создала свою версию RSS (1.0), близкую к исходной версии RSS 0.90 и максимально приближенную к стандарту RDF. RSS 1.0 предоставляет больше возможностей чем все 0.9x , например, допускает расширение при помощи модулей. Компания же UserLand решила развить ветвь 0.9x и создала версии 0.92, потом 0.93, 0.94, которые позволяют представлять метаданные, и наконец 2.0. При этом RSS 2.0 – не новая версия RSS 1.0, а логическое продолжение ветви 0.9x. В ней также добавлена поддержка модулей. В настоящее время существует 7 независимых версий RSS - RSS 0.90, 0.91, 0.92, 0.93, 0.94, 1.0, 2.0. Эти версии отличаются друг от друга, хотя все они ориентированы на один тип информации и содержат одинаковые базовые поля. При этом, многие считают все версии, кроме 2.0, устаревшими и «отмененными», но это далеко не так, пока еще самой популярной является RSS 0.91. Что же касается верстн 0.94, то ее спецификация не сохранилась даже на авторском сайте Userland. Так, по адресу http://backend.userland.com/rss094 находится спецификация версии RSS 2.0. Спецификации отдельных версий формата RSS приведены на таких Web-страницах:

RSS 0.90: http://www.purplepages.ie/RSS/netscape/rss0.90.html

RSS 0.91: http://my.netscape.com/publish/formats/rss-spec-0.91.html

RSS 0.92: http://backend.userland.com/rss092

RSS 0.93: http://backend.userland.com/rss093

RSS 1.0: http://web.resource.org/rss/1.0/

RSS 2.0: http://backend.userland.com/rss/

Во всех версиях RSS есть некоторые особенности, но объединяет их ориентация на один тип информации, вследствие чего они содержат общие базовые поля: основной блок данных (channel), который содержит атрибуты заглавия канала (title), ссылки (link), данные о языке сообщений (language) и логотип (image), после которых идет список самих сообщений, где в каждом пункте (item) указывается заголовок (title), краткое описание (description) и ссылка на новость (link). Кроме того, каждый RSS-файл начинается обязательными элементами xml и rss. Первый из этих элементов содержит атрибуты version (версия) и encoding (кодировка).

Среди множества необязательных элементов RSS можно назвать самые распространенные - язык (language), copyright, категория информации (category), дата и время публикации сообщения (pubDate), программа, которая использовалась для создания файла (generator), картинка, которую следует показывать наряду с текстовой информацией (image).

Кроме заголовка блока данных в формате RSS предусмотрено описание отдельных информационных элементов (item). Каждый элемент <item> - это отдельная статья или краткая аннотация и ссылка на полную версию статьи. Канал (channel) может содержать любое число элементов <item>, содержащих только два обязательных вложенных элемента - название (title) и описание (description). Кроме того, часто используются такие вложенные элементы: ссылка на первоисточник (link), категория (category), комментарий (comments) и автор (author).

В качестве примера новостного канала формата RSS 0.91 можно привести динамический файл, формируемый по адресу http://online.infostream.ua/rss.php (Обзор основных событий дня “Електронні Вісті”), имеющий такой вид:

```
<?xml version="1.0" encoding="windows-1251" ?>
<!DOCTYPE rss PUBLIC "-//Netscape Communications//DTD RSS 0.91//EN"
"http://my.netscape.com/publish/formats/rss-0.91.dtd">
<rss version="0.91">
<channel>
<title>Електронні Вісті</title>
```



```
<language>ru</language><image>
<title>Електронні Вісті</title>
<url>http://www.elvisti.com/images/export/elvisticom3_88x31.gif</url>
<link>http://www.elvisti.com</link>
<width>88</width>
<height>31</height>
</image>

<item><title>РАДАР СЛЕДИТ ЗА КОСМИЧЕСКИМ МУСОРОМ</title>
<description>В японской префектуре Окаяма с 6 апреля начал работать радар
с дистанционным управлением, основная функция которого состоит в
отслеживании перемещения космического мусора.</description>
<link>http://elvisti.com/2004/04/06/sci-tech.shtml#3</link>
</item>

<item><title>В ИВАНО-ФРАНКОВСКОЙ ОБЛАСТИ КУРИЦА СНЕСЛА ЯЙЦО ВЕСОМ 143
Г</title>
<description>В селе Делиев Галицкого района Ивано-Франковской области курица
снесла яйцо весом 143 г. </description>
<link>http://elvisti.com/2004/04/06/misc.shtml</link>
</item>

<item><title>В США БОЛЕЕ 60% КОРПОРАЦИЙ В 1990-Е ГОДЫ НЕ ПЛАТИЛИ
НАЛОГИ</title>
<description>Более 60% американских корпораций в период бума
американской экономики с 1996 по 2000 годы не платили налоги в
государственную казну, сообщило Главное бюджетно-контрольное
управление США.</description>
<link>http://elvisti.com/2004/04/06/biz.shtml#2</link>
</item>

<item><title>СЕДЬМОЕ АПРЕЛЯ - ВСЕМИРНЫЙ ДЕНЬ ЗДОРОВЬЯ</title>
<description>В нынешнем году по рекомендации ВОЗ этот день пройдет под
лозунгом "Безопасность на дорогах зависит от каждого из нас".</description>
<link>http://elvisti.com/2004/04/06/health.shtml#2</link>
</item>

</channel>
</rss>
```

Помимо формата RSS, недавно появился формат Atom 3.0 (http://www.mnot.net/drafts/draft-nottingham-atom-format-02.html), пока окончательно не утвержденный, но используемый на крупнейшем поисковом

портале Google, что предопределяет его популярность. Открытый стандарт Atom совершенствуется командой программистов из IBM, Google и других компаний.

Как и RSS, Atom является подмножеством XML. Приведем пример файла в формате, чтобы подчеркнуть его близость с RSS:

```
<?xml version="1.0" encoding="utf-8"?>
<feed version="0.3" xmlns="http://purl.org/atom/ns#">
  <title>Наименьший возможный фид в формате Atom 3.0</title>
  <link rel="alternate" type="text/html" href="http://diveintomark.org/"/>
  <modified>2004-04-09T18:30:02Z</modified>
  <author>
    <name>Иванов Петр</name>
  </author>
  <entry>
    <title>Atom 0.3 пример</title>
    <link rel="alternate" type="text/html"
     href="http://uaport.ua/2004/04/09/atom03"/>
    <id>tag:uaport.ua,2004:4.2397</id>
    <issued>2004-04-09T08:29:29-04:00</issued>
    <modified>2004-04-09T18:30:02Z</modified>
  </entry>
</feed>
```

Дэйв Уинер (Dave Winer), один из главных разработчиков RSS, недавно призвал разработчиков объединить свои усилия и разработать единый формат, совместимый как с RSS, так и с Atom, чтобы собрать конкурентные стандарты в единое целое. «Новый формат можно назвать RSS/Atom, — заявил Уинер. — Он бы имел всю функциональность, которую разработчики Atom обещают внедрить. Максимально авторитетный формат получил бы наиболее полную поддержку от всех разработчиков». Уинер предлагает, чтобы у RSS/Atom было как можно меньше отличий от RSS 2.0.

Еще один диалект XML - OPML (Outline Processor Markup Language) используется для описания совокупности RSS-фидов, спецификация которого размещена по адресу http://opml.scripting.com/spec. С помощью OPML обеспечивается эффективный унифицированный обмен списками RSS-фидов. Приведем фрагмент OPML-файла, обеспечивающий доступ к новостям службы “All Headline News” (http://www.allheadlinenews.com/feeds.opml):


```
<?xml version="1.0" encoding="ISO-8859-1" ?>
<opml version="1.0">
  <head>
    <title>All Headline News</title>
    <dateCreated>Mon, 12 Apr 2004 04:00:01 GMT</dateCreated>
    <dateModified>Mon, 12 Apr 2004 04:00:01 GMT</dateModified>
    <ownerName>AllHeadlineNews.com</ownerName>
    <ownerEmail>feeds@allheadlinenews.com</ownerEmail>
    <expansionState></expansionState>
    <vertScrollState>1</vertScrollState>
  </head>
  <body>
    <outline text="All Headline News">
      <outline text="All Headline News - Accounting"
htmlUrl="http://www.allheadlinenews.com/news/Accounting" language="en" title="All
Headline News - Accounting" type="rss" version="RSS"
xmlUrl="http://www.allheadlinenews.com/cgi-
bin/news/xml/newsxml.cgi?cat=Accounting" />
      <outline text="All Headline News - Acupuncture"
htmlUrl="http://www.allheadlinenews.com/news/Acupuncture" language="en" title="All
Headline News - Acupuncture" type="rss" version="RSS"
xmlUrl="http://www.allheadlinenews.com/cgi-
bin/news/xml/newsxml.cgi?cat=Acupuncture" />
      <outline text="All Headline News - Adolescent Health"
htmlUrl="http://www.allheadlinenews.com/news/Adolescent%20Health" language="en"
title="All Headline News - Adolescent Health" type="rss" version="RSS"
xmlUrl="http://www.allheadlinenews.com/cgi-
bin/news/xml/newsxml.cgi?cat=Adolescent%20Health" />
      <outline text="All Headline News - Adventure Sports"
htmlUrl="http://www.allheadlinenews.com/news/Adventure%20Sports" language="en"
title="All Headline News - Adventure Sports" type="rss" version="RSS"
xmlUrl="http://www.allheadlinenews.com/cgi-
bin/news/xml/newsxml.cgi?cat=Adventure%20Sports" />
    </outline>
  </body>
</opml>
```

Для доступа ко всем новостям службы “All Headline News” пользователю достаточно указать адрес http://www.allheadlinenews.com/feeds.opml в соответствующем окне своей программы чтения RSS, поддерживающей OPML

(например, FeedDemon). В списке доступных RSS-фидов сразу же окажутся более 100 каналов службы, таких как:

All Headline News – Accounting

All Headline News – Acupuncture

All Headline News - Adolescent Health

All Headline News - Adventure Sports

All Headline News – Advertising

All Headline News - Aerospace

Основным применением RSS в настоящее время являются новостные фиды (feed). Фид - это файл в формате RSS, в который записывается новостной контент Web-ресурса. Если есть необходимость оперативно отслеживать изменения на сайте, содержащем фид, то можно делать это с помощью программы-агрегатора, не посещая сайт непосредственно с помощью стандартных программ-браузеров.

Ниже приведены адреса самых популярных в Интернет фидов:

http://w.moreover.com/categories/ocs/ocsdirectory.rdf

http://10.am/extra/ocsdirectory.php

http://www.newsisfree.com/ocs/directory.xml

http://blogspace.com/rss/feeds/converted.ocs

http://www.groksoup.com/ocs/ocsdirectory.xml

http://theweb.startshere.net/channels.phtml?format=OCS

http://myrss.com/catalog/ocs04.rdf

http://www.syndic8.com/xml.php

В настоящее время в русскоязычной части Интернет представлены тысячи RSS-фидов, наиболее популярные из которых такие:

NEWSru.com – http://www.newsru.com/plain/rss/all.xml

Газета.ru - Все новости (RSS) (www.gazeta.ru/export/gazeta_rss.xml)

Lenty.RU - http://www.lenty.ru/export/bestnews.rss

Подробности – (http://www.podrobnosti.com.ua/export/)

Lenta.ru – (http://lenta.ru/l/r/EX/import.rss)

Полит.РУ – (http://www.polit.ru/rss/index.xml)

Портал "Юридическая Россия" (http://law.edu.ru/**rss**/news.**rss**)

Водка он-лайн - http://vodka.com.ua/export/rss.xml

Портал "ПлейМобайл" - http://playmobile.ru/news/rss

3Dnews - http://www.3dnews.ru/expnews/rss/newsrss.xml

Обширный список RSS-фидов русскоязычного сегмента Интернет находится по адресу http://my.yandex.ru/rss.opml; приведем лишь некоторые, наиболее интересные новостные фиды:

Аргументы и Факты - http://www.aif.ru/info/rss.php?magazine=aif

АвтоОБЗОР – http://auto.obzor.ru/news/autonews.xml

АвиаПорт.Ру – http://www.aviaport.ru/news/yandex_export.xml

Деловая Хроника – http://www.chronicle.ru/l/r/EX/rsschannel.xml

K2Kapital - http://ad.k2kapital.com/cbp/mynetscape/mynews.news

Linux.org.ru - http://images.linux.org.ru/getrss.php3

PalmQ Online - http://www.palmq.net/backend.php

СПОРТ сегодня - http://www.sports.ru/sports_docs.xml

TRAVEL.RU. Все о путешествиях - http://www.travel.ru/inc/side/yandex.rdf

АПК-Информ - http://www.apk-inform.com/yandextr.php

ФОНТАНКА.РУ - http://www.fontanka.ru/_transmission_for_yandex.thtml

IMA Press. Тема дня - http://www.ima-press.ru/rss.php?newsblock=theme

Журнал “Итоги” - http://www.itogi.ru/WebExport.nsf/Anons/itogi.xml

Остров. Новости Донбасса – http://www.ostro.org/yandex.php

ПОЛИТ.РУ - http://www.polit.ru/rss/index.xml?yandex_mode=1

PRAVDA.Ru - http://export.pravda.ru/yandex.txt

PR NEWS (все пресс-релизы компаний) –

http://www.prnews.ru/yandex/business.asp

Энциклопедия поисковых систем –http://www.searchengines.ru/news/news.rdf

Сетевой журнал - http://www.setevoi.ru/weekly/export1.txt

На сегодня существует уже множество служб синдикации новостей, которые предоставляют в доступ тематические фиды, построенные на основе использования многочисленных источников. Такой фид, к примеру, доступен на

портале UAport (http://uaport.net) и позволяет получить интегрированный доступ к потоку украинских и российских новостных сообщений, собираемому системой InfoStream®. С помощью RSS-шлюза системой InfoStream предоставляется унифицированный доступ к информации с 2500 Web-сайтов, сгруппированной по тематикам, языкам, странам, источникам. Объем этой информации сегодня превышает 40 000 сообщений в сутки. RSS-каналы UAport могут генерироваться системой по собственным запросам пользователей к поисковой системе.

Рассмотрим функциональность отдельных служб синдикации новостей, предоставляющих информацию в формате RSS.

Moreover

Для интеграции соответствующего запросам пользователей контента в корпоративные сети или порталы служба Moreover (http://www.moreover.com) использует собственное решение - Connected Intelligence. Прием информации в систему от 6500 источников в режиме реального времени осуществляется каждые 15 минут, классифицируется и группируется по темам.

На сайте Moreover содержатся сведения о технологических подходах к интеграции новостей, которые были созданы в этой службе и де-факто стали стандартами в системах мониторинга. Определена следующая технологическая цепочка: сначала выполняется оценка информационного содержания вэб-ресурса и построение конфигурационных профилей, описывающих данный ресурс. Редакторы в автоматизированном режиме оценивают ресурсы и формируют профили, удовлетворяющие информационным потребностям клиентов. Затем Web-ресурсы автоматически сканируются в соответствии с профилями, происходит преобразование информации в формат XML с добавлением RSS-тэгов. При этом устраняется дублирование. В соответствии с заданными правилами выполняется автоматическая классификация информации и загрузка ее в базы данных. Служба обработки запросов учитывает содержательную часть и требования к регламенту доставки. На последнем этапе происходит вывод и

доставка информации клиентам на их Web-сайты, в интранет-сети, на входы различных программных приложений.

В июле 2003 года технология Moreover была интегрирована в новостной портал Yahoo!, с сайта которого (http://news.yahoo.com) возможен доступ к информации из 3500 источников.

Google

В 2002 году популярная поисковая система Google запустила свой новостной сервис - Google News (http://news.google.com), который охватывает информацию с 4500 различных сайтов за последние 30 дней. Данные на сайте системы отсортированы по нескольким категориям, таким как международные новости, деловой мир, шоу-бизнес, технологии и спорт.

Новости в системе отбираются в зависимости от времени их публикации, популярности источника информации и количества статей, появившихся в Интернете на данную тему. Компания Google – популяризатор и один из разработчиков формата Atom, применяемого в основном в блогах.

Вместе с тем компания Google с подозрением относится к широким возможностям RSS-синдикации, видя в этой технологи возможности для нарушений авторских прав. Так, недавно Google запретила британскому веб-мастеру использовать результаты поиска в системе Google News на другом сайте в виде RSS-фида. Британский программист Джулиан Бонд создал скрипт на языке PHP, который берет введенный пользователем запрос, направляет его на Google News, а результат выдает в формате RSS. Полученный результат можно использовать в любом RSS-агрегаторе. Этот скрипт под названием gnews2rss можно найти на сайте http://www.voidstar.com/gnews2rss.php. По словам Бонда, основной протест со стороны Google вызвал не сам скрипт, а использование его для формирования новостной ленты на постороннем сайте. Скрипт все еще доступен в интернете и его можно использовать в программах-агрегаторах. Тем не менее в письме Бонду в Google указывали на то, что предпочтительным вариантом является применение службы Google News Alerts.

NewsIsFree

Одна из самых перспективных в Сети служб синдикации новостей NewsIsFree (http://www.newsisfree.com) охватывает свыше 9000 источников (в том числе российских и украинских). Сообщения обновляются каждые 15 минут и группируются по 15-ти основным категориям (http://www.newsisfree.com/sources/browse/). Примечательно, что режим поиска в RSS-ресурсах обеспечивается поисковым механизмом компании Google. Основная особенность службы NewsIsFree - это полная интеграция с XML, в частности, с RSS 0.91. Большинство разделов сайта службы содержат ссылки Syndicate, активизация которых приводит к отображению кода разделов в формате XML.

Несмотря на то, что основу информационных ресурсов, охватываемых службой, составляют англоязычные источники, NewsIsFree сегодня крупнейший интегратор и русскоязычных RSS-фидов, каталог которых доступен по адресу: http://newsisfree.com/sources/bylang/?lang=ru.

MSDN

Учитывая существующие в мире тенденции, служба MSDN (http://msdn.microsoft.com) также приступила к публикации своих новостей в формате RSS, выбрав версию 2.0. Ниже приведен список некоторых тем и адресов новостных фидов MSDN:

.NET Framework (http://msdn.microsoft.com/netframework/rss.xml)

ASP.NET - http://msdn.microsoft.com/asp.net/rss.xml

FrontPage - http://msdn.microsoft.com/office/frontpage/rss.xml

Longhorn - http://msdn.microsoft.com/longhorn/rss.xml

Mobile and Embedded - http://msdn.microsoft.com/mobility/rss.xml

MSDN Subscriptions - http://msdn.microsoft.com/subscriptions/rss.xml

Office - http://msdn.microsoft.com/office/rss.xml

Security - http://msdn.microsoft.com/security/rss.xml

Visual Basic - http://msdn.microsoft.com/vbasic/rss.xml

Visual C# - http://msdn.microsoft.com/vcsharp/rss.xml

Visual C++ - http://msdn.microsoft.com/visualc/rss.xml

Visual FoxPro - http://msdn.microsoft.com/vfoxpro/rss.xml

Visual J# - http://msdn.microsoft.com/vjsharp/rss.xml

Visual Studio - http://msdn.microsoft.com/vstudio/rss.xml

Web Services - http://msdn.microsoft.com/webservices/rss.xml

Windows Embedded - http://msdn.microsoft.com/embedded/rss.xml

Яндекс-Новости

Служба "Яндекс" открыла проект Яндекс.Новости (http://news.yandex.ru) , к которому в настоящее время присоединилось свыше 1400 Интернет-изданий. Новости сортируются по десяти категориям, существует возможность поиска новостей с указанием раздела и времени публикации новости. Поиск новостей возможен как по всем источникам, так и по заданным пользователем. Имеется также возможность поиска за произвольный период времени. Для сбора и экспорта новостей используется формат RSS 2.0.

Сегодня бесплатная служба синдикации новостного контента «Яндекс» представляет такие основные каналы:

Главные новости - http://news.yandex.ru/index.rss

Политика - http://news.yandex.ru/politics.rss

В мире - http://news.yandex.ru/world.rss

Общество - http://news.yandex.ru/society.rss

Экономика - http://news.yandex.ru/business.rss

Спорт - http://news.yandex.ru/sport.rss

Происшествия - http://news.yandex.ru/incident.rss

Культура - http://news.yandex.ru/culture.rss

Здоровье - http://news.yandex.ru/health.rss

Компьютеры - http://news.yandex.ru/computers.rss

Интернет - http://news.yandex.ru/internet.rss

Авто - http://news.yandex.ru/auto.rss

InfoStream

Разработанная в Информационном центре "ЭЛВИСТИ" система InfoStream® (http://infostream.ua) предназначена для автоматизированного сбора информации с открытых Web-сайтов, ее обработки, систематизации и обеспечения доступа к ней. Если пользователь хочет получать новостную информацию по интересующей тематике по e-mail, SMS или встроить постоянную подборку в свою веб-страницу, то к его услугам сервис InfoStream Client .

Персонализация интерфейса пользователей, работающих в режиме онлайн, т.е. сохранение их постоянных запросов и организация подписки, реализуется на основе современной технологии RSS 0.91. Для получения тематической ленты InfoStream (RSS-фида), в соответствующее поле RSS-агрегатора следует ввести адрес в формате:

http://uaport.net/UAnews/rss.php?<ЗАПРОС>,

где в качестве запроса можно ввести слово или словосочетание на языке запросов информационно-поисковой системы InfoReS.

На основе технологии InfoStream® пользователям доступны такие новостные каналы, сформированные по запросам аналитиков компании ElVisti:

Агропром - http://uaport.net/UAnews/rss.php?rubr01

Банки - http://uaport.net/UAnews/rss.php?rubr02

Экономика - http://uaport.net/UAnews/rss.php?rubr03

Экономика Украины - http://uaport.net/UAnews/rss.php?rubr04

Недвижимость - http://uaport.net/UAnews/rss.php?rubr05

Биржи - http://uaport.net/UAnews/rss.php?rubr06

Инвестиции - http://uaport.net/UAnews/rss.php?rubr07

Приватизация- - http://uaport.net/UAnews/rss.php?rubr08

Нормативные акты - http://uaport.net/UAnews/rss.php?rubr09

Оборона, Конверсия - http://uaport.net/UAnews/rss.php?rubr10

Официальная хроника - http://uaport.net/UAnews/rss.php?rubr11

Криминал - http://uaport.net/UAnews/rss.php?rubr12

Обзоры прессы - http://uaport.net/UAnews/rss.php?rubr13

Связь - http://uaport.net/UAnews/rss.php?rubr14

Экология - http://uaport.net/UAnews/rss.php?rubr15

Энергетика - http://uaport.net/UAnews/rss.php?rubr16

Медицина - http://uaport.net/UAnews/rss.php?rubr17

Наука и техника - http://uaport.net/UAnews/rss.php?rubr18

Компьютеры - http://uaport.net/UAnews/rss.php?rubr19

Астрология - http://uaport.net/UAnews/rss.php?rubr20

Культура - http://uaport.net/UAnews/rss.php?rubr21

Катастрофы - http://uaport.net/UAnews/rss.php?rubr22

Образование - http://uaport.net/UAnews/rss.php?rubr23

Внешнеэкономическая деятельность – http://uaport.net/UAnews/rss.php?rubr25

Масс-медиа - http://uaport.net/UAnews/rss.php?rubr26

Калейдоскоп - http://uaport.net/UAnews/rss.php?rubr27

Религия - http://uaport.net/UAnews/rss.php?rubr28

Спорт - http://uaport.net/UAnews/rss.php?rubr29

Туризм - http://uaport.net/UAnews/rss.php?rubr30

Автотранспорт - http://uaport.net/UAnews/rss.php?rubr32

Транспорт - http://uaport.net/UAnews/rss.php?rubr31

Для нахождения RSS-фидов существуют многочисленные списки и каталоги, однако объемы существующих RSS-ресурсов таковы, что пользователям уже не достаточно десятка-другого рубрик первого уровня, присутствующих в каталогах. Как всегда в подобных случаях, на помощь приходят информационно-поисковые системы, которые позволяют находить как целые RSS-фиды, так и отдельные сообщения по ключевым словам. Поэтому в Интернет появились поисковые сайты по RSS-фидам. Одним из первых был сервис Feedster.com, который кроме непосредственно поиска позволяет

подписаться на его результаты в формате RSS. В настоящее время Feedster обрабатывает 500 тысяч RSS-сообщений в сутки.

Еще одна поисковая система доступна на сайте http://Assimilatethe.net. Эта система охватывает свыше 3500 RSS-ресурсов. Система ищет по заголовкам и описаниям RSS-сообщений. В базе данных системы Assimilatethe сейчас порядка 193,000 сообщений.

Как известно, RSS – самый распространенный формат для “живих журналов” - блогов (от слова Weblog). Для поиска по блогам также существуют сотни каталогов и поисковых систем. Среди основних поисковых систем по блогам можно назвать:

DayPop – http://www.daypop.com

Blog Search Engine – http://blogsearchegine.com

Feedster – http://www.feedster.com

BlogStreet – http://www.blogstreet.com

Blogarama – http://blogarama.com/in.php?ID=2080

Globe of Blogs – http://www.globeofblogs.com

BlogDex – http://blogdex.media.mit.edu

Weblogs.com - http://weblogs.com

BlogWise – http://www.blogwise.com

BlogHop – http://www.bloghop.com

BlogUniverse – http://www.bloguniverse.com

Пользователи, конечно же, могут читать RSS-файлы с помощью стандартних Web-браузеров, что однако сопряжено с просмотром XML-разметки и полным отсутствием всякого оформления. За это и боролись создатели формата RSS. А вот для интерпретации этого формата существует бесчисленное множество программ, созданных в основном в последние два-три года. То есть, пользователи могут получить доступ к данным в формате RSS с помощью специальных программ. Эти программы называются RSS-агрегаторами и в наглядном виде отображают содержание RSS-фидов.

Программа-агрегатор позволяет собирать RSS-файлы с Web-сайтов, одновременно следить за появлением на них новостей и читать их содержание. Программы-аггрегаторы (их еще называют RSS-парсерами) выполняют синтаксический разбор данных, представленных в формате RSS, после чего могут реализовывать любые действия по отношению к этим данным, например, отсылать их по электронной почте либо отображать на определенном Web-сайте. Сегодня наиболее популярны агрегаторы, позволяющие собирать и группировать RSS-данные с разных Web-сайтов.

Feedreader (http://www.feedreader.com)

Feedreader - это свободно распространяемая программа для Windows, позволяющая читать данные в формате RSS версий 0.9, 0.91, 1.0, а также различную информацию от таких систем, как Dublin Core и Slashback (стандарты описания метаданных информационных ресурсов Сети). Утилита очень удобна в использовании, обеспечивает работу с информацией на русском и украинском языках и обладает широким кругом сервисных возможностей. FeedReader версии 2.5 доступен по адресу http://sourceforge.net/project/showfiles.php?group_id=70179.

Feedreader – типичный RSS-агрегатор, интерфейс котрого напоминает интерфейс почтовых программ. У пользователя, знакомого с почтовыми клиентами, работа с программой не вызывает затруднений. Остановимся подробнее на самых необходимых возможностях этой программы.

Для настройки подписки на RSS-фид пользователю следует активизировать опцию New и ввести следующую информацию:

- адрес RSS-фида;
- название фида (оно может быть определено пользователем);
- периодичность обращения к фиду на Web-сайте для обновления.

При этом имеется возможность изменения кодировки, размеров шрифтов, помещения фида в отдельную папку, группировки фидов. Для управления подпиской существуют дополнительные опции, активизируемые нажатием правой клавиши мыши при установке курсора на конкретном фиде:

- обновление фида (списка активных сообщений);
- отметка всех сообщений как уже прочитанных;
- удаление списка сообщений;
- изменение свойств подписки, включая тему, периодичность и др.

Для получения полного текста сообщения (на которое есть ссылка – <link>), заголовок и аннотация которого вызвали интерес, следует:

- произвести двойное нажатие левой клавиши мыши на заголовке, или
- нажать на ссылку "Read on" в поле аннотации, или
- нажать на соответствующую кнопку, стоящую перед заглавием, или
- нажать правую клавишу мыши, находясь курсором на заглавии, при этом можно открыть текст сообщения в новом окне браузера, или
- активизировать ссылку первоисточника и выйти через сеть Интернет на первоисточник.

FeedDemon (www.feeddemon.com)

FeedDemon представляет собой коммерческую программу, обеспечивающую удобную работу с RSS версии 2.0. Имеется возможность опробовать работу програмы в «триальном» режиме. Утилита работает в среде Windows, корректно обращается с русской и украинской кодировками, обеспечивает поиск-фильтрацию информации фидов. Триал-версия FeedDemon 1.0 находится по адресу http://www.feeddemon.com/download/dloadhandler.asp?file=feeddemon-trial.exe, размер инсталлятора – 2,3 МБайта. В дружественном пользователю интерфейсе агрегатора легко отслеживать и читать свежие фиды. FeedDemon позволяет представлять содержимое новостных лент в виде своеобразной газеты.

Приступить к использованию программы можно немедленно после инсталляции, так как сразу пользователь начнет получать рассылки с сайтов Rollingstone.com, Scripting News, Sladshot, Wired, Yahoo! и др. Сообщения программа позволяет сохранять (News Bins) и отслеживать по ключевым словам, запуская функцию Watches. Отдельные RSS-фиды можно перенаправлять в

тематические списки или каналы. FeedDemon также позволяет проводить поиск и читать новости в автономном режиме.

Для подписки на фиды в программе следует ввести URL источника или импортировать файл OPML.

Abilon и ActiveRefresh

Эти два агрегатора от одного производителя - компании Abilon и ActiveRefresh (http://www.activerefresh.com/download.php). Бесплатная программа Abilon вполне подходит для среднего пользователя, программа проста и надежна, отличается высокой скоростью и малой ресурсоемкостью (339 КБ). Она обладает возможностью закачки новых каналов с сайтов MoreOver, MyRss и NewsIsFree. Однако ей не хватает возможностей глобального поиска и сжатия информации.

В отличие от Abilon, ActiveRefresh - это платная программа - полная реализация концепции компании, которая позволяет агрегировать обычные Web-сайты, импортировать с них новости, представленные в HTML, следить за почтовыми ящиками, проводить глобальный поиск и т.д.

Syndirella 0.9b

Syndirella (Синдирелла) может показывать информацию как с обычных Web-страниц, так и отображать данные, представленные в формате RSS. Программа реализована на платформе .NET, функционирует в среде операционных систем Windows и требует установки Internet Explorer версии 5.0 или выше. Для работы программы необходимо инсталлировать библиотеку Microsoft .NET Framework runtime версии 1.0 (20 Мб). Однако если эта компонента уже установлена, то сама программа Syndirella займет всего 250 Кб. Адрес для загрузки: http://www.yole.ru/projects/syndirella.

Сегодня большую популярность, кроме перечисленных, для работы под Windows получили еще два агрегатора - Awasu и Beaver. Особенность бесплатной программы Awasu (http://www.awasu.com) заключается в ее возможности объединять потоки множества новостных сайтов и блогов. Beaver

(http://www31/brinkster.com/toolmaker) принимает фиды форматов RSS/RDF и имеет привычный интерфейс в стиле Outlook Express.

K.R.S.S. 2.6

KDE's Rich Site Summary viewer - приложение для Linux, позволяющее отображать данные в формате RSS на экране в виде HTML-страниц. Есть возможности по настройке вида отображения при помощи Cascading Style Sheets (CSS) и установки специальных фильтров новостей. Адрес для загрузки программы: http://krss.sourceforge.net/downloads.html , размер файла - 394 Кб.

Liferea

В последнее время для ОС Linux большую популярность приобретает агрегатор Liferea (http://liferea.sourceforge.net/). Liferea поддерживает многочисленные основанные на XML форматы новостных фидов, такие как RSS, RDF, Atom, Echo, PIE, а также OCS и OPML для списков фидов. Эта программа распространяется с библиотекой GTK2.

В настоящее время создаются и уже созданы многочисленные инструментальные средства для разработки программ работы с RSS-данными. Например, для разработки программ-парсеров на языке Perl создан модуль XML::RSS, который загружается с сайта http://search.cpan.org/.

Встраиваемые в Internet Explorer инструментальные полосы (тулбары) от Dogpile (http://www.dogpile.com/info.dogpl/tbar/) и HotBot Desktop (http://www.hotbot.com/tools/desktop/) поддерживают технологии RSS и Atom. С помощью этой возможности заголовки сайтов, поддерживающих RSS, просматриваются прямо не выходя из браузера.

Одна из самых заметных черт интерфейса будущей весии ОС Windows - Longhorn заключается в наличии многофункциональной боковой панели (Sidebar). На нее может быть помещена любая информация - от часов и списка контактов до новостей, импортируемых в формате RSS. При этом средства настройки панели включены в состав инструментария разработчиков и поддаются настройке с их

стороны.

Владельцы карманных компьютеров, установив на свои устройства RSS-агрегаторы, могут эффективно просматривать новостные файлы в RSS - формате. Для платформы Palm OS наиболее популярной является программа компании Stand Alone - Hand RSS. Стоит эта программа $14.95, но скачать и опробовать ее демо-версию можно бесплатно (http://standalone.com/palmos/hand_rss/).

В качестве еще одного эффективного агрегатора можно назвать программу Quick Palm RSS Reader (http://remus.manilasites.com/).

Из специализированных для Pocket PC можно назвать агрегатор новостей в RSS/RDF PocketFeed (http://www.furrygoat.com/Software/). Пятнадцатидневная демо-версия еще одной программы для этой платформы (PocketPC 2002 и Windows Mobile 2003) - PocketRSS 1.3 доступна на сайте http://www.happyjackroad.com/AtomicDB/pocketpc/pocketRSS/pocketRSS.asp.

Не обязательно устанавливать программу-агрегатор прямо на наладоннике. Существуют серверные решения, выполняющие всю работу по интерпретации RSS-фидов и преобразованию результатов к формату, пригодному для КПК. Один из лучших сайтов подобного назначения - MobileRSS (mobilerss.net). Для работы с этим бесплатным сервером необходима лишь формальная авторизация. Зарегистрированный клиент вводит и активизирует адреса необходимых ему RSS-фидов, после чего просматривает их в свободном режиме. Примечательно, что этот зарубежный сервис обеспечивает корректную работу с кириллическими шрифтами.

С помощью современной RSS-технологии пользователи Интернет получили надежный и простой доступ к ресурсам оперативной информации с Web-сайтов Сети. Перспективность и популярность RSS как стандарта обусловлена прежде всего его доступностью и простотой. Сегодня практически все ведущие информационные сайты в мире, "живые журналы", работающие в Интернет, используют RSS как инструмент оперативного представления обновлений своих ресурсов.

Еще один, неожиданный аспект применения RSS-технологий стал актуален в связи с массовым распространением невостребованных рассылок по электронной почте – СПАМа. Действительно, электронная почта привлекательна и для спамеров. Нередко списки электронных адресов подписчиков новостей на сайтах и порталах становятся добычей взломщиков. Этот фактор делает подписку через e-mail достаточно рискованным занятием. Поэтому можно предположить, что на смену рассылкам придет использование RSS-фидов. В отличие от рассылок по электронной почте, где доставка инициируется администраторами сайтов, после того как подписчик оставил им свой адрес, в случае с RSS пользователь сам вводит адрес необходимого ему RSS-фида в программу-агрегатор. Эта программа периодически проверяет, не изменилось ли содержание RSS-фида, и при наличии изменений автоматически закачивает его содержимое. Главным преимуществом RSS-технологии оказалось то, что пользователь сам принимает решение о получении каждого конкретного сообщения.

Популярность RSS-технологии у владельцев Web-ресурсов набирает все большую популярность еще и благодаря своей экономичности – не требуется никаких средств борьбы со спамом, фильтрации писем, управления рассылкой. При этом все, кому это необходимо, получают требуемую информацию о важных событиях, корпоративных анонсах, обновлениях Web-сайтов.

4. Инфраструктура информационных прокси-серверов

Возможности доступа к Интернет-ресурсам, которые привлекают своей открытостью, объемами и содержательной многогранностью, на первый взгляд кажутся безграничными. Однако кризисные события в разных областях – будь то крупные теракты или чемпионаты по футболу, свидетельствуют об обратном. Именно в кризисных ситуациях Интернет достаточно часто подводит. Существует множество проблем – от перегруженности сетевой инфраструктуры – до вирусных атак, уязвимостях и отказов в обслуживании отдельных Web-серверов. Целый ряд проблем порожден также объемами, разнообразием представления и динамикой контентной части сетевых информационных потоков.

Таким образом, несмотря на такие позитивные качества, как открытость и доступность, существующую инфраструктуру Интернет нельзя признать надежной, живучей и достоверной [7]. Назовем еще несколько проблем, присущих современному Web-пространству:

1. Не решена задача доступа пользователей к разнородным Web-ресурсам «из одного окна» для получения обобщенного представления потоков информации по необходимой тематике.
2. Не обеспечена возможность своевременного "напоминания" и "проталкивания" профильной для пользователя информации, публикуемой на большом количестве Web-сайтов.
3. Достаточно высокая вероятность отказа в обслуживании со стороны критически важных Интернет-ресурсов в самое необходимое время.

Известно, что сегодня существуют технологии интеграции контента, частично предоставляющие решение названных проблем, однако не исследован уровень безопасности их применения, возможно массового. Вопросы сетевой безопасности, например, в рамках современной концепции Семантического Web, по мнению авторов, выглядят преимущественно декларативно, а на практике заужены тематикой цифровой подписи.

Из сказанного выше следует необходимость создания новой инфраструктуры, обеспечивающей надежную доставку сетевого контента заинтересованным лицам и организациям, в частности, на государственном уровне.

Пожалуй, самая распространенная причина отказов от предоставления Web-сайтами своего контента по запросам пользователей состоит в их банальной перегруженности. Вместе с тем мало кто из информационных администраторов Web-сайтов, даже сайтов и порталов органов государственной власти, владеют данными о максимально возможном количестве запросов пользователей, которые способны удовлетворить эти ресурсы. Владельцы любительских Web-сайтов и сайтов электронных СМИ даже не задумываются об этом.

При этом существуют достаточно жесткие ограничения возможностей Web-сайтов при массовой работе с их контентом. Следует заметить, что многие из этих ограничений не учтены даже в нормативных документах, регламентирующих требования по защите информации на Web-страницах. Назовем некоторые из них, которые влияют на уровень доступности Web-ресурсов:

- ширина канала связи до Web-сайта. Это ограничение было наиболее обосновано на начальных этапах развития Интернет;
- физические ограничения программно-технических платформ Web-серверов. Для снятия этого ограничения, например, популярные поисковые службы используют сотни Frontend-серверов;
- устанавливаемые ограничения в программном обеспечении Web-серверов. Например, у самого популярного в настоящее время Web-сервера Apache [34] параметром MaxKeepAliveRequests определяется максимальное количество разрешенных запросов при устойчивом соединении. При этом для обеспечения максимальной производительности это значение зачастую устанавливается по умолчанию равным 100;
- ограничения на отдачу динамических страниц, например, со стороны СУБД, поисковых систем или сервисных других программ. Такие ограничения часто устанавливаются при совместном виртуальном

хостинге у провайдеров и измеряются количеством запросов в час. В случае использования популярной в Интернет СУБД MySQL [9] это ограничение, например, задается параметром max_questions, значение которого, как правило, составляет 72000 (20 обращений к базе данных в секунду). Превышение ограничения может происходить по разным причинам: установка малого значения в соответствии с политикой провайдера, высокая посещаемость сайта, установка ресурсоемких приложений типа статистики, нестандартных программ и т.д.

Следует выделить два явления, которые существенно влияют на надежность получения информации от Web-сайтов: пиковые нагрузки со стороны пользователей в кризисные дни (например, 11 сентября, «Оранжевая революция», начало войны в Ираке и т.п.) [36] и DoS-атаки (Denial of Service или Отказ от обслуживания). Во втором случае хакеры особым образом формируют запросы к программным компонентам Web-серверов, чтобы загрузить их до такого уровня, когда они перестанут функционировать. Такие атаки, как правило, не ведут к разрушению самих серверов, чтобы вернуть Web-сервер в рабочее состояние, как правило, требуется перезагрузка. Часто DoS-атака выполняется с большого количества компьютеров, в этом случае она называется распределенной (DDoS Distributed Denial of Service). Этот вид атак можно отнести к так называемым «сетевым войнам», формам организации конфликтных ситуаций на основе Интернет. В таких случаях Web-серверы не успевают отвечать на все запросы, в том числе и запросы реальных пользователей.

Обе ситуации – и злонамеренная DoS-атака, и кризисная пиковая посещаемость приводят к недоступности информационных ресурсов Web-сайтов, в частности, для аналитиков и лиц, принимающих решения.

Поведение систем в результате возникновения данных ситуаций: определенное количество запросов может обрабатываться – остальные стоят в очереди или «отбрасываются» по тайм-ауту.

Как подход к решению названных проблем предлагается построение сети – системы связанных информационных прокси-серверов. Необходимо заметить,

что использование прокси-серверов (точнее, кэширующих прокси-серверов) при работе в Интернет очень популярно [14]. В этом случае прокси-серверы служат в основном для ускорения загрузки страниц за счет кэширования содержимого страниц, ответов на запросы пользователей, DNS и т.п.

Для английского слова “proxy” в данном контексте применимы такие переводы: «полномочный представитель», «посредник». В Интернет-технологиях прокси – это программа, которая получает запросы, обращается к внешнему сервису из Интернет, получает ответы и возвращает их пользователям. Под кэшем понимается информационное хранилище, в котором хранятся часто запрашиваемые Web-страницы.

Именно идеологию кэширующего прокси-сервера предлагается рассмотреть как базу для построения инфраструктуры, которая позволит решить названные проблемы.

При этом к данным, которые предположительно будет обслуживать информационный прокси-сервер, предъявляются такие требования:

- рассматривается динамическая новостная составляющая web-пространства, как наиболее критичная с точки зрения обеспечения оперативного доступа;
- множество кэшируемых Web-сайтов выбирается экспертами в соответствии с вкладом этих источников в информационное пространство и может ограничиваться несколькими тысячами;
- информация в прокси-сервере должна быть представлена в универсальном внутрисистемном формате, предполагающем однозначную синтаксическую трактовку. Этим форматом может быть популярный сегодня XML или один из его диалектов (например, RSS);
- данные в информационном хранилище (кэше) должны обновляться и ротироваться по расписанию, соответствующему динамике их обновления на Web-сайтах.

Прокси-сервер, с одной стороны, предназначен для надежного обслуживания пользователей корпоративных сетей, а с другой стороны, может

обеспечивать обмен данными с аналогичными внешними прокси-серверами. Такое взаимодействие образует своеобразную сетевую структуру, которая, по мнению авторов, может оказаться решением названных проблем.

Пользователи информационного прокси-сервера обращаются к данным, помещаемым в информационное хранилище (кэш). Кэш пополняется программой-роботом, которая сканирует целевые Web-сайты. Следует отметить, что многие популярные сетевые информационно-поисковые системы также кэшируют информацию с Web-страниц, предоставляя ее при необходимости пользователям. Можно назвать такие системы, как Yandex (режим «Сохраненная копия»), Rambler (режим «Восстановить текст»), Google (режим Cached).

Характерная особенность роботов - настойчивость (при получении отказов на запросы, он продолжает их задавать до момента получения позитивного ответа). Это тот плюс, который, например, позволил авторам наблюдать поток сообщений из Вашингтона 11 сентября при общем впечатлении об Интернет, как «зависшей» в тот момент Сети.

Интеллектуальный сканер системы (рис. 33) обращается к Web-сайтам и скачивает с них информацию по сценарию, составленному на специальном языке макроописаний [20]. При этом сценарии могут существенно отличаться по качеству, все зависит от квалификации эксперта-оператора.

Предполагается, что в результате сбора и первичной обработки данные в информационном хранилище будут программно приведены к единому формату, классифицированы в соответствии с определенными рубрикаторами, каждому документу приписан ряд дескрипторов, включая ключевые слова.

Вместе с тем администраторам Web-сайтов известны многие роботы, которые излишне загружают их ресурсы, не принося при этом явной пользы.

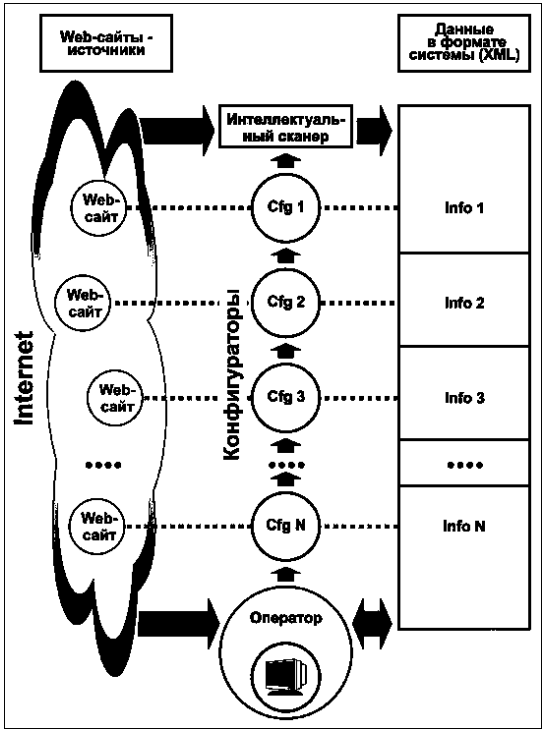


Рис. 33. Процедура сбора данных

Опасность массового применения роботов состоит в том, что они сами могут порождать нечто подобное DoS-атакам. Что можно противопоставить этой опасности? По мнению авторов, это:

- строгое соблюдение стандарта исключений для роботов (этот документ можно найти, например, по адресу http://www.robotstxt.org/wc/exclusion.html);
- аккуратное описание сценариев сбора информации роботами, зачастую буквально эмуляция действий пользователей;
- создание сети информационных прокси-серверов, например, на отраслевых уровнях. В этом случае сканироваться могут не Web-сайты-оригиналы, а ближайшие прокси-серверы.

На рис. 34 приведен принцип функционирования сети информационных прокси-серверов. На нем представлен иерархический принцип организации этой сети. Прокси–сервер первого уровня обеспечивает доступ к кэшу, заполняемому интеллектуальным сканером. К этому кэшу с помощью информационно-поисковой системы обеспечивается доступ конечных пользователей корпоративной сети. Эти же пользователи имеют возможность обращения к документам непосредственно в Интернет. Представленные на рис. 34 прокси-серверы второго уровня загружают информацию с кэша прокси-сервера 1-го уровня, а кроме того, могут дополнять свое информационное хранилище данными, сканируемыми непосредственно из Интернет (информационные потребности пользователей разных прокси-серверов могут отличаться). Очвидно расширение приведенной схемы на третий и последующие уровни.

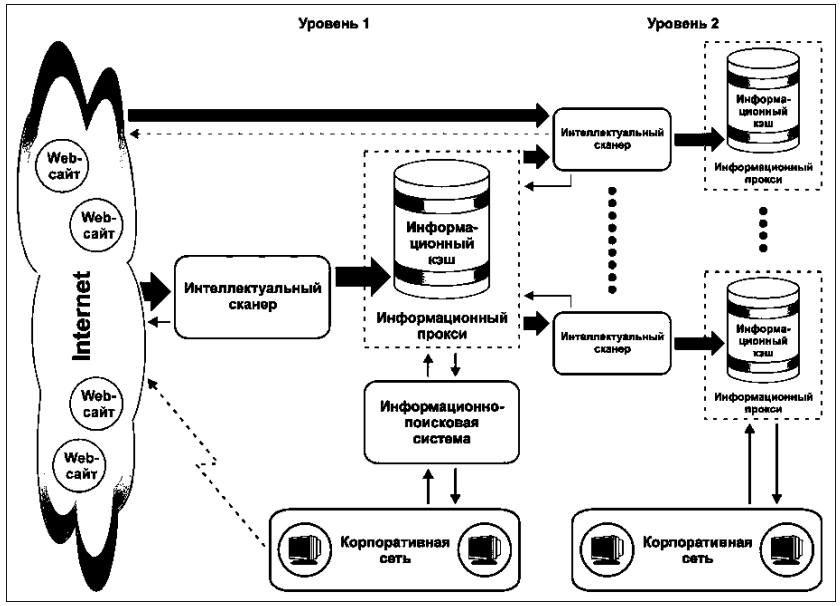


Рис. 34. Принцип организации сети информационных Proxy-серверов

В качестве прототипа информационного прокси-сервера рассматривается система, созданная на основе комплекса мониторинга новостей InfoStream [9], которая в настоящее время позволяет осуществлять сканирование информации из нескольких тысяч открытых Web-сайтов.

На основе этой системы реализуетися информационный прокси-сервер, к которому обращаются пользователи – корпоративные серверы, которые сами непосредственно не сканируют Интернет (или выполняют эту операцию в ограниченных объемах, решая специфические информационные задачи). Такой подход обладает следующими преимуществами:

1. Не требуется сканирования и обработки данных из Интернет непосредственно (прежде всего - экономия на ресурсах, необходимых для администрирования).
2. Анонимность (при сканировании сайтов их владельцы могут определять адреса робота-сканера).
3. Существенная экономия Интернет-трафика (в этом случае основные расходы берет на себя информационный провайдер - владелец первого проски-сервера. Как показывает опыт, соотношение объемов сканируемой и «готовой к употреблению» информации составляет 50:1).
4. Не отрицается возможность самостоятельного сканирования Интернет (например, ресурсы общего плана можно загружать из информационного прокси-сервера, а специальные ресурсы – непосредственно из Интернет).

Для корпоративных пользователей реализовано решение InfoStream Port, которое обеспечивает доступ к базам данных оперативной и ретроспективной информации в корпоративных сетях. Программно-технологическое обеспечение InfoStream Port основано на принципе интеграции информационного прокси-сервера и поисковой системы, и включает как компоненты утилиту обмена данными с информационным хранилищем (кэшем) и полнотекстовую информационно-поисковую систему InfoRes.

Информационное обеспечение системы у корпоративного пользователя, функционирование которой основывается на использовании кэша, формируется

за счет выполнения совокупности технологических операций, в число которых входят сбор информации из Интернет, нормализация информации, приведение ее к единому системному формату, классификация, помещение данных в информационное хранилище и предоставление санкционированного доступа к кэшу.

Описанная распределенная система информационных прокси-серверов позволяет создавать эффективные и масштабируемые решения, которые могут быть существенным подспорьем для аналитиков, сотрудников информационных служб, так как они способны существенно повысить надежность доставки и уровень обобщения оперативных данных, а также снизить загрузку каналов связи. Благодаря используемому кэшированию не только повышается эффективность использования каналов, но и уменьшаются задержки, возникающие в процессе доставки интернет-контента пользователю.

Критически важным в этой технологии являются инструментальные средства, которые должны гарантировать безопасность, актуальность принимаемых и передаваемых данных, а также их целостность.

5. Проблема дублирования информации

Сегодня Интернет-пространству присущи такие недостатки, как ограниченность интегрированного доступа к информационным ресурсам, обилие «информационного мусора», невозможность гарантирования целостности документов, практическое отсутствие возможности смыслового поиска [1]. Эти проблемы обуславливаются несколькими причинами, среди которых можно назвать непропорциональный рост уровня информационного шума и многократное дублирование информации.

Важные сообщения многократно дублируются на экспоненциально растущем количестве сайтов, в то время как количество заслуживающих внимания источников растет не такими высокими темпами, скорее всего, линейно. Дело в том, что серьезные источники информации - это объекты реальной жизни, в то время как сайты в своей совокупности представляют виртуальное пространство, которое развивается по собственным законам.

Задача выявления дублирующихся сообщений (их принято называть дубликатами), а также перепечаток документов с небольшими изменениями («почти дублей») является одной из актуальнейших и сложнейших при интеграции информационных ресурсов. Понятие содержательных дублей документов достаточно расплывчато, авторы даже пытались анализировать такие явления, как пересказ одних и тех же событий, описание различных аспектов разными людьми.

В свое время определенные надежды возлагались на развитие т. н. семантических методов, которые бы позволили оперировать непосредственно со смыслом сообщений, и таким образом избежать проблем его формализации. Однако они не оправдались.

С прагматической точки зрения применения таких методов следует выделить два главных недостатка. Это существенная зависимость практической реализации метода от языка обрабатываемых дкументов (что фактически делает невозможной работу с многоязычными потоками) и его неустойчивость: для

некоторых информационных массивов (вероятно, подобных тем, на которых данная система настраивалась) результаты очень хорошие, но для других – очень плохие.

Пессимистический взгляд на применение «семантических» методов в области информационных технологий, в общем-то, вполне понятен в чисто теоретическом плане. Действительно, семантика занимается отношением лингвистических конструкций к предметам и явлениям окружающего нас реального мира, тогда как компьютерные системы могут манипулировать исключительно формальными элементами. Иными словами, в рамках любой информационной технологии машина может устанавливать отношения только одних лингвистических конструкций с другими лингвистическими же конструкциями. Вопрос о том, в какой мере все это может имитировать семантические связи, остается открытым.

С другой стороны, вообще игнорировать семантические аспекты информационных технологий, несомненно, было бы ошибкой. Во всяком случае, интуиция и опыт подсказывают, что понятие семантической близости документов должно иметь определенный смысл и на уровне машинной обработки текстов.

Серьезное упрощение названой задачи может быть получено за счет применения содержательных методов, например, путями ранжирования первоисточников, определения и выделения тематических информационных каналов, экспертного формирования словарей значимых слов и т.п.

Преодоление использования явно дублирующейся информации не представляет проблем, однако дублирующиеся по смыслу сообщения выявляются не так легко, здесь на помощь приходят алгоритмы, аналогичные алгоритмам построения информационных портретов [4], их сопоставления и вероятностной оценки. На практике явные дубликаты выявляются даже с помощью механизмов контрольных сумм, но этот подход не решает проблем пользователей, для которых чаще всего не имеет значения, с чем они имеют дело, с прямой перепечаткой или с небольшой перефразировкой. Вместе с тем многие недобросовестные издания перепечатывают содержание сообщений, попросту

изменяя заглавия (работа хедлайнеров). И такой вид дублирования элементарно обходится с помощью контрольных сумм (но уже без учета заголовков). Дальнейший анализ показал, что при перепечатке материалов чаще всего остаются без изменений несколько первых предложений текста или первый абзац. И этот критерий был учтен и успешно внедрен. Вместе с тем качество выявления содержательного дублирования оставалось недостаточно высоким.

Исследовались методы, основанные на учете повторений встречаемости цепочек слов, например, метод «шинглов» (чешуек), достаточно хорошо описанный в работах [57], [60], [43] и [44]. Этот остроумный и эффективный метод поиска «почти дублей» оказался не очень чувствительным для небольших текстов с возможными перефразировками (авторы с интересом наблюдали эффекты двойного перевода при перепечатках с русского на украинский, а затем снова на русский).

В рамках предлагаемой модели авторы с самого начала отказались от попыток в полной мере понять, как именно происходит семантическая интерпретация текста. Предполагалось лишь то, что ключевую роль здесь играет принципиальная возможность представить текст в виде множества слов.

Наиболее прямой путь к установлению связи между произвольным документом и семантическим пространством предполагает наличие некоторого морфизма между устойчивыми сочетаниями слов и единицами смысла. При всей своей внешней банальности, это утверждение отнюдь не тривиально, поскольку речь в нем идет именно о морфизме, но отнюдь не об эквивалентности.

Устойчивое сочетание слов само по себе вовсе не является единицей смысла. Более того, далеко не всегда единица смысла вообще может быть артикулирована с помощью набора слов. Но между наборами слов и единицами смысла всегда или почти всегда могут быть установлены (вообще говоря, неоднозначно) устойчивые отношения.

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 [30]. При этом качество слов определялось экспертами, а соответствующий математический аппарат получил название «нечеткой цифровой сигнатуры». В этом подходе авторов смутил наивный подход, например, при умножении вероятностей зависимых событий (слов в сообщениях), а также необходимость «ручного» отбора значимых слов (очевидно, важность отдельных слов может изменяться во времени).

Изначально в распоряжении авторов был достаточно мощный информационный ресурс одной из служб интеграции новостей - ретроспективная база данных системы контент-мониторинга InfoStream [18]. Следует отметить, что процент дублирующихся сообщений в системе InfoStream значительно меньше, чем во всем Интернет-пространстве. Это объясняется подбором источников для сканирования, в число которых входят лишь те, которые хоть изредка публикуют оригинальные материалы.

Обработка входных сообщений в системе контент-мониторинга InfoStream, вплоть до выявления значимых ключевых слов, представлена на рис. 35.

Принцип выявления значимых ключевых слов (далее будем называть их *термами*) базируется на законе Зипфа [30], [71] и сводится к выбору слов со средней частотой встречаемости (наиболее встречаемые слова игнорируются с помощью «стоп-словаря», а редкие слова из текстов сообщений не учитываются).

В качестве некоторых «инвариантов» для отдельных сообщений в системе InfoStream используются цепочки из 12 наиболее весомых с точки зрения лингвостатистических критериев термов, 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).

Итак, выявление дублирующихся по содержанию новостных сообщений в системе InfoStream выполняется на основе лингвостатистических методов, заключающихся в выявлении в различных документах общих термов, цепочки которых образуют словесные сигнатуры сообщений.

Рис. 35. Схема обработки входных сообщений

Метод, предложенный авторами и используемый в системе InfoStream, заключается в признании документов дубликатами, если в их сигнатурах совпадает более 5 термов (из 12 возможных). Следует отметить, что применение более «мягкого» критерия к множеству отобранных термов позволяет реализовать режим «поиска подобных документов».

Введем обозначения: пусть "$\prec$" – оператор подобия, а "$\equiv$" - оператор дублирования. Очевидно, что для алгоритма выявления подобных документов и дубликатов, о котором идет речь, справедливо правило рефлексивности:

$$A \prec A, \quad A \equiv A,$$

где A – произвольный документ.

Оператор подобия не обладает свойством симметричности. Из подобия документа A документу B не следует обратное, т.е.:

$$A \prec B \not\Rightarrow \quad B \prec A.$$

Также не выполняется условие транзитивности:

$$A \not\prec B, \quad B \prec C \not\Rightarrow \quad A \prec C.$$

Действительно, например, отельный документ может быть подобен тексту из подборки, которая его включает, но сама подборка может не быть подобной

этому документу. Или документ может быть подобен двум документам, из которых он скомпилирован, но сами оригиналы могут существенно отличаться.

Для отношения дублирования, наоборот, симметричность и транзитивность выполняются:

$$A \equiv B \Rightarrow \quad B \equiv A,$$

$$A \equiv B, \quad B \equiv C \Rightarrow \quad A \equiv C.$$

Заметим, что отношение, обладающее свойствами рефлексивности, симметричности и транзитивности является отношением эквивалентности [39], в нашем случае, отношением содержательного совпадения или дублирования.

Как было замечено, свойство дублирования документов является более жестким критерием подобия, например, совпадение 3, 4 или 5 термов свидетельствуют о некоторой содержательной близости, т.е. можно записать:

"$\prec$" $\rightarrow$ "$\approx$".

На практике каждому документу D_i из контрольного документального корпуса по приведенному выше алгоритму совпадения термов в сигнатурах (в разных экспериментах варьировались необходимые количества совпадающих термов) ставился в соответствие вектор с элементами:

$$a_{ij} = \begin{cases} 1, & D_i \equiv D_j, \\ 0, & \text{иначе.} \end{cases}$$

Условие симметричности в этих обозначениях записывается следующим образом:

$$\forall \quad i, \quad j: \quad a_{ij} = a_{ji},$$

а условие транзитивности:

$$\forall\ i,\ j,\ k: \quad a_{ij} = 1, \quad a_{jk} = 1 \Rightarrow a_{ik} = 1.$$

Авторы исследовали критерии подобия (изменяя количество сравниваемых в сигнатурах термов), чтобы достичь на контрольном документальном корпусе максимального уменьшения коэффициента асимметричности:

$$\frac{\sum_{i}^{N}\sum_{j}^{N}|a_{ij}-a_{ji}|}{\sum_{i}^{N}\sum_{j}^{N}a_{ij}},$$

и увеличения коэффициента транзитивности:

$$\frac{\sum_{i}^{N}\sum_{j}^{N}\sum_{k}^{N}a_{ij}a_{jk}a_{ik}}{\sum_{i}^{N}\sum_{j}^{N}a_{ij}},$$

где N – количество документов в контрольном корпусе.

Очевидно, что так рассчитываемый коэффициент асимметричности ассоциируется с огрублениями при определении дубликатов, а уровень транзитивности – с полнотой.

Вместе с тем следует заметить, что проверка коэффициентов асимметричности и транзитивности может использоваться лишь для формальной проверки приближения отношения к свойствам эквивалентности. Само определение того, что эта эквивалентность – содержательное дублирование было предоставлено аналитиками-экспертами. Приведенный выше алгоритм, кроме своего эмпирического подтверждения, хорош тем, что позволяет варьировать некоторым числом (количеством сравниваемых термов в сигнатурах), значение которого можно подобрать с учетом оптимизации двух названных коэффициентов.

На рис. 36 и 37 приведены экспериментально полученные значения коэффициентов симметричности и транзитивности в зависимости от количества учитываемых совпадающих термов при попытках выявления дублирования. При определении оптимального количества термов, необходимого для выявления дублирования учитывался баланс этих коэффициентов. Кроме того, параллельно результаты оценивались экспертами. Следовало бы отдельно остановиться на субъективном факторе, присутствующем при экспертной оценке уровня

дублирования. Сегодня учет этого фактора такая же сложная и неоднозначная задача, как и задача определения пертинентности результатов поиска [30].

Опыт показал, что в русско- и украиноязычных потоках новостей совпадение хотя бы 6 термов в сигнатурах документов приводит к более чем 90% полноты и 95% точности при выявлении содержательных дубликатов.

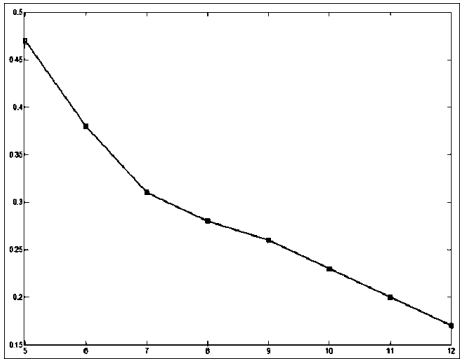


Рис. 36. Зависимость коэффициента асимметричности от количества совпадающих термов в критерии выявления дубликатов

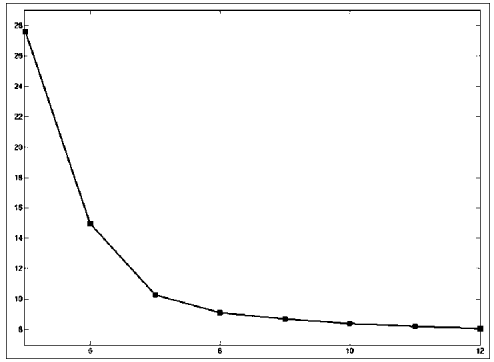


Рис. 37. Зависимость коэффициента транзитивности от количества совпадающих термов в критерии выявления дубликатов

В соответствии с этим критерием, авторами было проведено исследование соотношения дублирующихся и оригинальных сообщений в новостных

информационных потоках. Исследования привели к удивительному результату. Оказалось, что количество оригинальных сообщений и их содержательных дублей, охватываемых системой InfoStream в 2005 году, почти в точности совпало (рис. 38).

Это же соотношение справедливо для отдельных событий, отражаемых в электронных СМИ (рис. 39). Лишь некоторые «феноменальные» публикации, дублируются десятки раз.

Авторами также исследовался уровень дублирования для новостных документов, имеющих контекстные ссылки на другие сайты-источники.

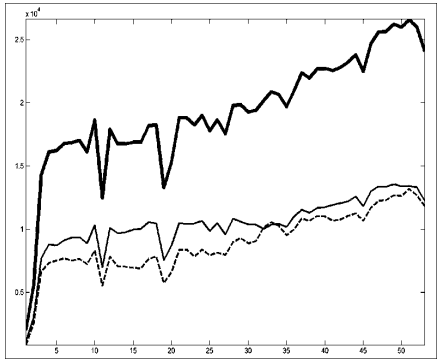


Рис. 38. Объемы информации, сканируемой системой InfoStream в 2005 году, в разрезе недель:
сплошная жирная линия – общий объем сообщений, сплошна тонкая – оригинальные сообщения, пунктирная – информационные дубли

На рис. 40 приведен график зависимости уровня дублирования для источников (исследовалось около 1500 сайтов), ранжированных по количеству исходящих ссылок. По графику видно, что до определенного значения (порядка 800) уровень дублирования значительно превышает средний, равный ~ 50%. При небольшом количестве исходящих ссылок этот уровень понижается, однако при минимальном количестве ссылок снова возрастает. Можно считать, что значения

рангов источников 1400 и выше соответствуют «зоне массового плагиата» (ссылок мало, а уровень дублирования - высокий).

Проведенные исследования позволили на новом уровне реализовать информационное обслуживание пользователей системы контент-мониторинга InfoStream, обеспечивая селекцию дубликатов. Кроме того, авторами был составлен список наиболее оригинальных информационных источников, сканируемых системой InfoStream, который представляет безусловный интерес для корпоративных пользователей.

Наряду с вышесказанным, необходимо 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 сообщение многократно дублируется на сайтах и в СМИ) или при определении эффективности PR-кампаний (подсчет републикаций пресс-релизов и др.)

Полученные результаты позволили вплотную подойти к решению проблемы эффективного автоматизированного выявления плагиата в текстах небольших объемов. Эта проблема сегодня имеет большой резонанс [29], [67], но существующие алгоритмы ее решения раскрываются не часто из-за опасений обесценивания наработанных механизмов.

В заключение следует назвать две проблемные области в выявлении дубликатов по представленному алгоритму. Во-первых – это некорректная во многих случаях работа с короткими сообщениями, зачастую вырождающимися в один лишь заголовок. Выявление значимых слов в таких сообщениях - проблема, не решенная авторами до сих пор.

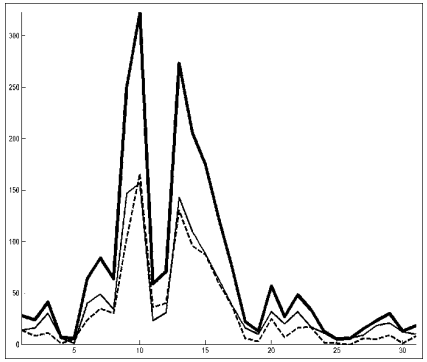


Рис. 39. Объемы информации, сканируемой системой InfoStream в марте 2005 года по запросу «Cebit».

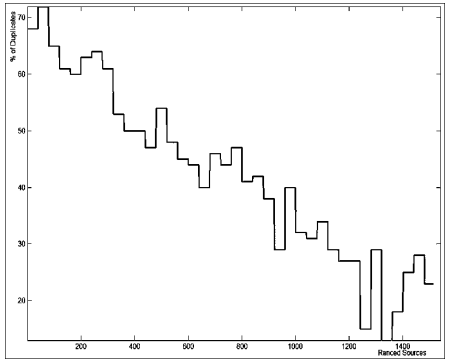


Рис. 40. Уровень дублирования в зависимости от ранга источников

Так, два сообщения с результатами футбольного матча на 10-й и 40-й минутах могут практически не отличаться по набору слов, разница будет лишь в счете. Вторая проблема связана с длинными документами, обзорами, дайджестами. Термы в словесных сигнатурах таких документов могут не отражать контента каждой составляющей обобщенного документа. Обе названные проблемы случая ведут к понижению полноты и точности при выявлении дубликатов и могут рассматриваться как открытая тема для дальнейших исследований.

6. Концепция аннотированного поиска

Вероятно, если вычленить центральную проблему современных информационных потоков, то она состоит в качественном различии понятий «релевантность» и «пертинентность». Сам факт наличия этих двух терминов говорит о том, что различие было известно всегда, но в условиях ограниченных объемов данных им можно было пренебречь, так как потребитель, в явном виде просмотрев всю релевантную выборку, мог отобрать то, что ему нужно.

Сегодня же, когда зачастую это становится невозможным, несовпадение релевантности с пертинентностью выступает на первый план. Действительно, если из 10 тысяч предъявленных информационно-поисковой системой (ИПС) документов все являются пертинентными, то потребитель будет удовлетворен, по крайней мере в первом приближении, прочитав любое их количество. Остальные он может просто проигнорировать без особого ущерба для достижения поставленной цели. В некоторых областях отмеченная закономерность эффективно используется. Так, например, службы синдикации новостей обслуживают своих клиентов, при том, что количество охватываемых источников информации практически у любой из них в настоящее время не превышает 10 тысяч. При этом следует отметить, что проблему полноты новостной информации такой подход позволил решить, оставив, однако, нерешенной проблему формирования достаточного для пользователя объема информации.

Существующие ИПС изначально проектировались для обеспечения именно релевантности выборки по отношению к формальным запросам, и в этом их главный недостаток в современных условиях. Низкий, а точнее говоря, неконтролируемый уровень пертинентности выборки с высоким уровнем ее релевантности порождает различные ситуации, допускающие более или менее общую типизацию.

Наиболее простой и очевидный случай: по запросу «президент» можно получить, кроме необходимых новостей, рекламу отеля «Президент», прайс-листы сигарет «Президент» и т. п. Теоретически, с таким информационным

мусором можно бороться, составляя уточняющие запросы из 200-300 поисковых терминов с активным использованием операторов контекстной близости и операций отрицания, но это сложная работа, требующая времени, определенной подготовки и практического опыта. Во всяком случае, у обычного пользователя есть шанс получить желаемое, прибегнув к помощи профессионалов.

Предположим, пользователя интересуют специальные работы по методам кодирования текстовой информации. Он составляет соответствующий запрос и получает набор документов, которые действительно посвящены этой теме. Но ему предъявляются классические учебники теории связи, тогда как требуется получить последние публикации с оригинальными результатами. В этом случае расширение запроса, скорее всего, не поможет, поскольку его обработка, какой бы сложной она ни была, предполагает использование того, что содержится в тексте документа в явном виде и может быть реализовано лингвистическими средствами.

Далее, пусть пользователь действительно получил ссылку на обзор по теме, содержащий то, что ему нужно. Но неприятность при этом заключается в том, что оказывается, что именно этот обзор уже лежит у него на столе, и, значит, нет нужды искать его в базах данных, а другие обзоры, возможно, находятся где-то в конце списка ссылок, но этого пользователь никогда не узнает. И справиться с такой ситуацией намного сложнее, чем с первыми двумя, потому что факт наличия у пользователя некоторых данных никак не отражен в самих данных.

Вряд ли есть смысл отдельно говорить о том, что сетевое информационное пространство в принципе структурируемо достаточно слабо. Более того, эволюция как Сети в целом, так и отдельных ее сегментов может служить некоторым примером стохастического процесса. Возможно, именно это обстоятельство и является главной причиной низкой эффективности организации быстрого прямого доступа к информационным единицам, о которых мы часто даже не знаем.

Сказанное не означает, что сетевое пространство является полностью хаотическим и может быть описано в терминах теории шума. На самом деле, оно

содержит в себе элементы упорядоченности (назовем их кластерами), в известной мере аналогичные доменам ферромагнетиков. Но их много, и каждый из них обладает собственной динамикой развития, слабо коррелирующей с другими такими динамиками. С другой стороны, кластеры могут интенсивно взаимодействовать друг с другом, порождая своего рода «отраженные волны» и формируя тем самым разнообразные обратные связи, как положительные, так и отрицательные.

Но, во-первых, сами кластеры не всегда являются устойчивыми во времени – они возникают, исчезают, меняют свои контуры, мигрируют и т. п., и, во-вторых, взаимодействие между ними носит вполне стохастический характер.

Первый реальный шаг к решению проблемы структуризации сетевого информационного пространства, очевидно, должен состоять в формировании некоего порожденного пространства, обладающего достаточным уровнем упорядоченности и в разумном приближении адекватного исходному. Таким образом, может быть поставлена задача, понимаемая как некоторое неоднозначное отображение неупорядоченного множества составляющих элементов сетевого информационного пространства на упорядоченное множество их образов, обладающее требуемой (например, иерархической) организацией.

Тогда поиск в широком смысле слова может производиться на структурированном множестве образов информационных единиц, а предъявление его результатов должно включать в себя восстановление оригиналов.

Конечно, приходится считаться с возможностью утраты части информации, но ведь и в стандартной реализации невозможно добиться одновременно релевантности результата и ее разумной полноты. Поэтому, так или иначе, мы можем лишь говорить о некотором ожидании получения требуемого, связанного с вероятностью его обнаружения.

По крайней мере, для определенного класса задач такой подход, в числе прочего, может решить до сих пор открытую проблему теории поиска – проблему информационного дублирования. Именно при построении пространства образов

могут формироваться цепочки более или менее информационно-подобных элементов, отображаемые затем на один и тот же образ.

Естественно, в процессе построения пространства образов, они могут снабжаться наборами метаданных.

Первой попыткой практического решения названной проблемы являются рубрицированные каталоги Web-сайтов, однако ее следует считать ограниченной по двум причинам: во-первых, как правило, классифицируются только сайты, а не входящие в их состав документы, а во-вторых, используется стандартный (и практически не зависящий от времени) набор предопределенных рубрик. Суть проблемы даже не в том, что рубрик слишком мало для полноценной структуризации сетевого информационного пространства, а в том, что они отражают не его реальные свойства, а субъективные представления потребителей о структуре предметной области. Так, например, внешняя статическая рубрикация не в состоянии локализовать реально существующие в данный конкретный момент кластеры.

По мнению авторов, структуризация сетевого информационного пространства неизбежно должна предполагать постоянное (фоновое) сканирование информационных потоков и создание их виртуального образа, предназначенного для практического использования. В идеале, этим могли бы заниматься специализированные службы, предоставляя результаты своей деятельности в распоряжение поисковых систем.

Одним из наиболее естественных путей решения подобных проблем нам представляется перенесение центра тяжести с наборов данных, в которых следует вести поиск, на ассоциируемые с ними наборы метаданных, содержащих широкий спектр внешних характеристик, по которым потребитель может достаточно просто сформировать своего рода «словесный портрет» требуемых документов. Ядро формируемого пользователем поискового предписания должно представлять собой набор формальных параметров, указывающих на выбор той или иной категории из содержащихся в метаданных. Традиционный же запрос,

включающий в себя ряд поисковых терминов, может играть здесь вспомогательную роль для сужения объема (теперь уже пертинентной) выборки.

Наборы метаданных могут, разумеется, создаваться специальными программными комплексами, входящими в состав поисковых систем, путем автоматической обработки сканируемой информации. Однако значительно большей эффективности можно было бы достичь, организовав хранение метаданных (пусть даже «сырых») непосредственно в информационных документах.

Для этих целей как нельзя лучше подходят XML-технологии, получившие в последнее время широкое распространение. Так, например, в состав Web-сайта мог бы входить XML-документ, содержащий некое унифицированное (в идеале – стандартизированное) описание структуры и основных предопределенных характеристик этого сайта, предназначенных для построения различного рода классификаторов.

Очевидно, что даже в простейшем случае подключения набора классификаторов мы получаем значительный выигрыш в объеме результатов поиска. Пусть мы имеем три классификатора, каждый из которых содержит десять категорий. В случае равномерного распределения документов по категориям получим фактор 10^{-3}, т. е. вместо 10000 документов потребитель получит их всего 10.

Разумеется, приведенные рассуждения касаются не только поиска в чистом виде, но и сопряженных задач, таких как, скажем, избирательное распространение информации.

Парадокс в развитии сетевых поисковых систем состоит в том, что их техническое совершенствование в рамках традиционной парадигмы неизбежно приводит к лавинообразному увеличению баз данных, и соответственно, объемов релевантных выборок, которые конечный потребитель в итоге не в состоянии обработать [1].

Существующие информационно-поисковые системы изначально проектировались для обеспечения релевантности выборки в сочетании с

требованием полноты поиска, но именно в этом и состоит их главный недостаток. Неконтролируемый уровень пертинентности выборки при этом резко снижает вероятность получения пользователем имено той информации, что ему требуется.

Причины избыточности результатов стандартного информационного поиска могут быть разделены на две качественно различные категории: дублирование информации и информационное несоответствие. Существенным является то, что принадлежность документа к числу дублей носит вполне объективный характер и может определяться автоматически на основании формальных критериев.

Напротив, информационное несоответствие порождает проблемы чисто субъективного характера, так как машина не в состоянии определить, соответствует ли содержание данного документа информационным потребностям данного пользователя.

Поэтому становится ясно, что поисковые технологии должны быть расширены за счет применения дополнительных семантических средств, позволяющих либо сократить разрыв между уровнями релевантности и пертинентности, либо как-то его компенсировать.

Наиболее перспективным из существующих сегодня направлений, несомненно, является автоматическое группирование результатов поиска [4], т. е. разбиение релевантной выборки документов на кластеры. Вместе с тем она не решает проблему по существу, поскольку хотя и помогает ориентироваться в результатах поиска, но отнюдь не приводит к сокращению их объемов.

Главное достоинство такого автоматического группирования состоит в иерархической организации результатов поиска, позволяющей на первом этапе иметь дело с ограниченным набором кластеров, а затем уже переходить к составу того или иного кластера. Сложность, однако, заключается в том, что разбиение выборки на группы осуществляется на основании формально понимаемой близости документов. Это обстоятельство, естественно, приводит к тому, что конечный эффект зависит от многих, в том числе и случайных, факторов и носит явно неконтролируемый характер.

Особую актуальность приобретают подходы, позволяющие переформулировать задачу поиска таким образом, чтобы его результаты действительно могли быть без труда восприняты пользователем.

Одним из основных принципов, положенных в основу более адекватных подходов, на наш взгляд, является отказ от требования полноты поиска.

Вполне разумной представляется постановка задачи предварительной обработки исходной совокупности документов, имеющей целью сформировать некоторый эффективный набор данных, отражающий в разумном приближении ее содержание и предназначенный для дальнейшего поиска по нему.

Сама по себе такая постановка задачи отнюдь не является новой: она широко и успешно применяется в сфере автоматического реферирования документальных потоков. Именно продуктивность подобной методики в смежной области и заставляет нас внимательно присмотреться к ее возможностям применительно к информационному поиску.

В технологическом плане предлагается реализация принципа предварительной обработки текстового материала с помощью методик, характерных для другой области информационных технологий, а именно контент-анализа. Такая обработка предполагает автоматическое выделение наиболее значимой информации и отсев «мусора», что позволит пользователю работать с наборами данных, достаточно ограниченными по объему, и, при правильной организации, может существенно повысить уровень пертинентности результатов поиска. Концепция также предусматривает своего рода кластеризацию, однако распределению по группам подлежит не только релевантная выборка, но и исходный набор документов, в котором ведется поиск.

В рамках концепции используются термины «аннотированный поиск» и «аннотированная база данных», поскольку, как будет видно ниже, основные алгоритмы поиска и структура базы данных напоминают те, которые используются при автоматическом реферировании.

Центральная идея предлагаемой концепции состоит в том, что релевантность документа следует определять по отношению к некоторому его

информационному образу. Причем последний должен быть построен именно так, чтобы отражать основное содержание документа. Такой образ документа (или группы документов) в рамках данной концепции называется аннотацией.

Структура и форма аннотации не имеют принципиального значения, но в любом случае она должна содержать упорядоченный набор терминов и/или фраз, входящих в состав соответствующего документа и обладающих определенным уровнем весовых значений. Вес может характеризовать значимость терминов или фраз в документе и может определяться различными методами в зависимости от свойств предметной области и поставленной задачи. Кроме того, поскольку потребителя в конечном счете интересуют тексты документов, совокупность аннотаций должна быть дополнена системой соответствующих ссылок. Вместе они образуют некий набор метаданных, который должен быть включен в общую индексную систему базы данных.

В качестве информационно-технологической основы рассматривается база данных традиционной информационно-поисковой системы с присущей ей структурой, включая, например, индексные, инверсные, словарные таблицы и т.п.

Создание аннотированной базы данных подразумевает создание базы данных поисковых образов первичных документов и их кластеризацию, т.е. автоматическое формирование групп документов с близкими по некоторым критериям поисковыми образами (ПОД).

При формировании аннотированной базы данных важнейший аспект – формирование базы данных аннотаций, т.е. поисковых образов кластеров (ПОК), которые, собственно, и будут использоваться в процессе поиска. Естественно, эта база данных связана с базой данных кластеров, каждая запись которой соответствует определенному кластеру и включает, кроме всего прочего, его описание (выполненное методами автоматического реферирования).

Методы автоматического реферирования (а точнее квазиреферирования, основанного на преимущественном использовании методов статистического анализа текстов) используются, с одной стороны, для создания ПОД, а с другой стороны и описаний, доступных пользователям.

Задача полнотекстового поиска по сверхбольшим текстовым массивам может оказаться неэффективными, например, в "Войне и мире" Л.Толстого можно найти большинство лексем русского языка. Поиск по аннотированным текстам в таких случаях решает проблему точности. Таким образом, вместо поиска по полным текстам оказывается целесообразным вести поиск по аннотациям - поисковым образам документов. Хотя квазиреферат часто для больших текстов оказывается образованием, лишь отдаленно напоминающим исходный текст, при этом зачастую не воспринимаемым человеком, но именно как поисковый образ документов, содержащий взвешенные ключевые слова и фразы, он может приводить к вполне адекватным результатам при полнотекстовом поиске.

Квазиреферат в большинстве известных систем строится из текстовых фрагментов, имеющих наибольшие весовые значения. Общий вес текстового блока на этом этапе определяется по формуле [18]:

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 резюмирующие конструкции-маркеры, типа "в заключение", "в данной статье", "согласно результатам 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 весов входящих в него терминов - слов и словосочетаний. После выявления определенного, заданного коэффициентом необходимого сжатия, количества текстовых блоков с наивысшими весовыми коэффициентами, они объединяются для построения квазиреферата.

Следует отметить, что не только аннотации в виде ПОК, но и описания отдельных элементов в базе данных аннотаций, доступной на этапе поиска,

создаются на основе средств автоматического реферирования, которые на этом этапе не учитывают предпочтений пользователей, выраженные поисковыми предписаниями (запросами).

На рис. 41 приведена схема функционирования аннотированной базы данных.

В рамках данной концепции предполагается использование методов квазиреферирования, преимущество которых заключается в простоте реализации.

При обращении пользователей к базе данных предполагается следующая процедура: запрос пользователя отрабатывается по базе данных аннотаций, после чего поисковой процедурой выполняется формирование набора релевантных кластеров, наименования и описания которых, с одной стороны, могут предъявляться пользователям (если их количество не превышает заданного заранее), а с другой стороны, если количество результатов поиска (кластеров) превышает это значение, то результаты поиска автоматически группируются, образуя суперкластеры, перечень которых и предъявляется.

Естественно, в последнем случае пользователю предъявляются названия суперкластеров и их описания – рефераты, составленные автоматически уже с учетом предпочтений пользователей. Вес текстовых фрагментов в этом случае описывается уточненной формулой:

Weight = Location + KeyPhrase + StatTerm + UserPref

Коэффициент *UserPref* - пользовательские предпочтения - зависит от того, насколько слова и словосочетания, приведенные в запросе пользователя, присутствуют в данном фрагменте.

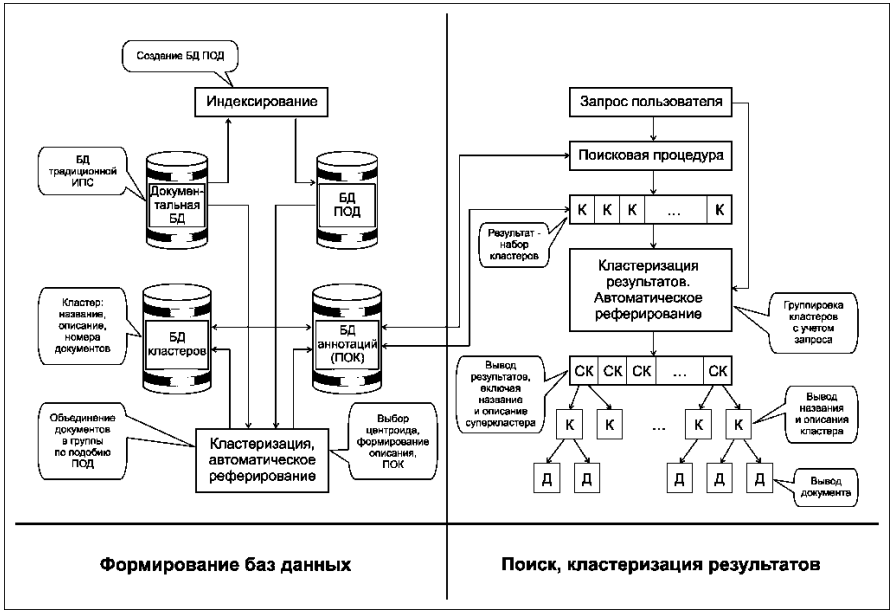


Рис. 41. Архитектура и модель функционирования аннотированной базы данных

Представление результатов поиска может осуществляться различными способами, в зависимости от особенностей предметной области, структуры документальной базы данных, характера информационных потребностей пользователей и т. д. Отметим лишь, что сами аннотации, как уже указывалось выше, являются поисковыми образами - внутренними элементами системы и пользователю в исходном виде не предъявляются. Поэтому предполагается, что в целях адекватного отображения результатов поиска каждый построенный кластер снабжается описанием, которое также строится автоматически и выдается пользователю как «этикетка» кластера, которая, в отличие от аннотации, представляет собой связный текст. Далее пользователь, если пожелает, может просмотреть все документы, входящие в состав данного кластера.

Предполагается, что при такой организации поиска релевантными окажутся лишь те документы, для которых поисковые термины запроса пользователя являются информационно-значимыми. Это достигается уже в силу того обстоятельства, что сами аннотации по своей природе обладают именно таким

свойством. Наличие в них исключительно терминов или фраз с достаточно большими весовыми значениями препятствует попаданию в релевантную выборку документов, в которых поисковые термины присутствуют в виде информационного шума.

Следует отметить, что приведенная модель в настоящее время еще не реализована полностью в виде программно-технологического обеспечения, однако отдельные элементы уже созданы и прошли достаточно большую апробацию. Осталось дело за малым - запустить эту модель на реальных сверхбольших объемах данных. К реализованным элементам относятся: традиционные полнотекстовые информационно-поисковые системы, включая авторскую разработку – систему InfoRes; алгоритмы автоматического реферирования; механизмы кластеризации как статических, так и динамических массивов информации, которые находят уже сегодня применение, например, при выявлении основных сюжетов в системе контент-мониторинга InfoStream; адаптивные интерфейсы уточнения запросов к информационно-поисковой системе.

Представленная модель ориентирована на практическую реализацию и в явном виде содержит ряд технологических ограничений, главное среди которых связано с тем, что на этапе индексирования поисковые образы документов создаются без учета предпочтений пользователей. ПОД не является полной копией документов, поэтому заранее не могут быть учтены все нюансы информационных потребностей пользователей, что может сказаться не только на полноте, но и на релевантности. Сгладить названную проблему могут лишь изощренные интеллектуальные алгоритмы автоматического реферирования.

Вместе с тем предлагаемая организация поиска позволит решить следующие важные задачи:

- автоматическое группирование документов и тем самым сокращение реального объема пространства поиска;
- предъявление пользователю исключительно информационно значимых документов;

- при необходимости исключение дублей из результатов поиска при сохранении их в самой базе данных.

Вспомним, что средняя длина запроса к поисковой системе в Интернет не превышает 2-3 слов, может быть в том числе и поэтому основные проблемы пользователя сводятся к разрешению проблемы релевантности-полноты, и, в конечном счете, пертинентности выдачи. Очевидно, предлагаемая система организации поиска позволит существенно повысить его привлекательность с точки зрения «среднестатистического» пользователя.

Так как, ввиду растущих объемов информации, поисковые системы уже сегодня не в состоянии предоставить в распоряжение все то, что требуется пользователю из текстового корпуса Интернет, реализация данной концепции даже на первом этапе поиска даст ему относительно небольшую выборку, позаботившись о том, чтобы она была содержательной.

7. Выявление новых событий

Рост объемов информации и скорости ее распространения фактически породил понятие информационных потоков [1]. Исследование такой составляющей этих потоков, как сообщений, публикуемых на страницах Web-сайтов, должно использовать принципиально новый инструментарий, так как классическй математический аппарат и инструментальные средства не всегда способны адекватно отражать ситуацию. В этом случае речь идет не столько об анализе документального массива фиксированного размера, сколько о навигации в потоке документов.

Несомненным является тот факт, что информационные потоки большей частью порождаются событиями реального мира. Действительно, возникновение информационных потоков можно представить себе как генерацию и движение наборов данных, ассоциированных с определенным событием, реализуемым как некоторый смысловой блок. Конечно, одному событию может соответствовать произвольное число сообщений. Таким образом, характеристики информационных потоков изначально определяются потоками событий реального мира.

С другой стороны, если событие новое и важное, то обязательно о нем будут много говорить в дальнейшем, то есть задача выявления новых событий из потока новостей является задачей предсказания дальнейшего появления множества «подобных» сообщений, задача прогноза [22].

Оптимальное решение, способное помочь ориентироваться в динамической части Интернет, сегодня предоставляют системы синдикации новостей [18], [36]. Под синдикацией в данном случае понимается сбор информации в Интернет и последующее распространение ее фрагментов в соответствии с потребностями пользователей.

Технология синдикации Интернет-новостей (рис. 42) включает в себя "обучение" программ сбора информации структурным особенностям отдельных

источников (Web-сайтов), непосредственное сканирование информации, ее приведение к общему формату (как правило, XML), а также классификацию.

Средства классификации и распределения информации представляют собой информационно-поисковую систему избирательного распространения информации (Информационный роутер). Документы, поступающие в систему, анализируются на соответствие тематическим запросам. Релевантные документы рассылаются пользователям, а также загружаются в тематические базы данных.

В режиме диалогового доступа к базам данных обеспечивается просмотр, поиск и отображение данных, а также предоставляется возможность обращения к оригиналам документов в Интернет.

Перспективным направлением развития технологии контент-мониторинга является глубинный анализ текстов (Text Mining), средствами которого обеспечивается решение задач формирования тематических информационных каналов, дайджестов, таблиц взаимосвязей и гистограмм распределения понятий. К этому классу средств можно отнести также выявление новых событий, их отслеживание и группировку (кластеризацию).

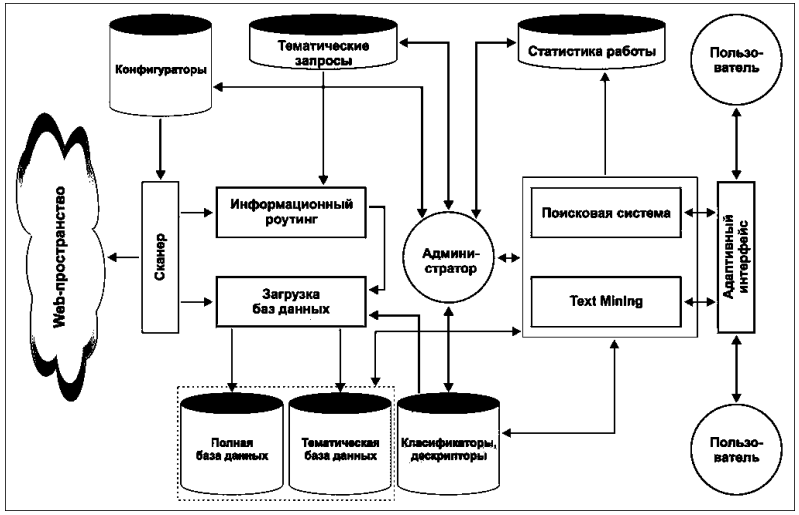


Рис. 42. Функциональная схема системы контент-мониторинга новостей из Интернет

Как правило, задача выявления новых событий из потока сообщений предполагает, что на вход соответствующего программно-технологического комплекса последовательно поступают новые документы. Они могут поступать как непосредственно от средств сканирования (политематический поток), так и от информационного роутера, системы избирательного распространения информации, отобранные по тематическому запросу (рис. 43). Далее, в соответствии с определенными алгоритмами, некоторые из которых приведены ниже, происходит выявление новых событий. 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), которые предположительно заполнятся в дальнейшем. В этом предположении и заключается прогнозный момент технологии выявления новых событий. Со временем каждый из кластеров может стать основой формирования полноценной сюжетной цепочки.

Алгоритм выявления основных сюжетных цепочек, используемый, например, в системе контент-мониторинга InfoStream [17], заключается в следующем. Последний поступивший на вход системы документ (документ с номером 1 при обратной нумерации) порождает первый кластер и сравнивается со всеми предыдущими. Если мера близости для какого-нибудь документа оказывается ближе заданной пороговой, то текущий документ приписывается первому кластеру. Сравнение продолжается, пока не исчерпывается список актуальных документов потока. После такой обработки документа 1, происходит обработка следующего документа, не вошедшего в первый кластер, с которым последовательно сравниваются все актуальные документы потока и т.д. В результате формируется некоторое неизвестное заранее количество кластеров, которые ранжируются по своим весам, задаваемым суммой нормированных метрик близости для всех элементов кластера.

Несмотря на то, что минимальный кластер может включать всего 1 документ, на окончательное рассмотрение принимается лишь определенное количество кластеров с наибольшими весами, т.е. группы наиболее цитируемых и актуальных сообщений. Для выбранных кластеров заново пересчитываются центроиды – документы, наиболее отражающие тематику кластера. Таким образом, формируются сюжетные цепочки, реализующие запросы типа «о чем пишут больше всего в последнее время?»

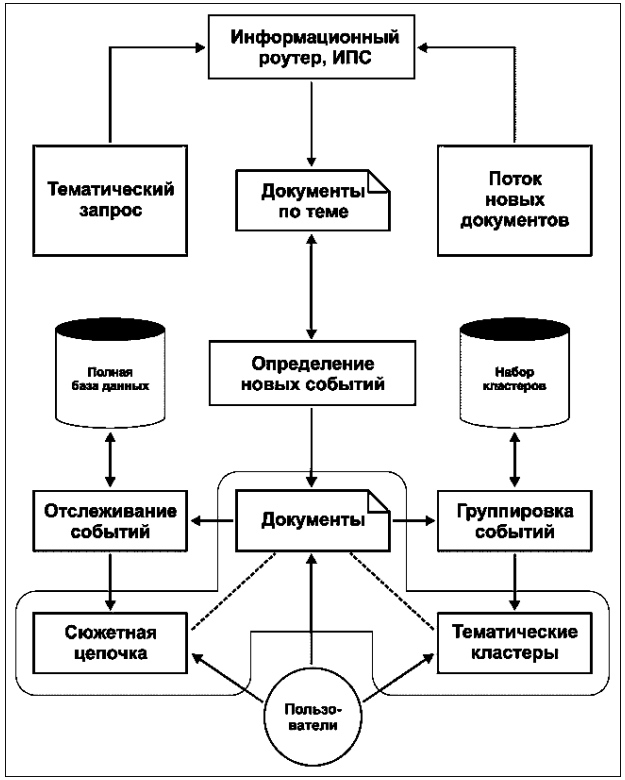


Рис. 43. Определение новых событий – одна из функций системы контент-мониторинга

При построении сюжетных цепочек система определяет лингво-статистические характеристики отобранных в результате поиска документов и автоматически выявляет наиболее значимые темы, освещаемые в

информационных потоках. Все весомые сообщения группируются по принадлежности к автоматически определяемым сюжетам. В качестве названий сюжетных цепочек используются заголовки сообщений, наиболее точно отражающих их суть. Порядок отображения сюжетов определяется количеством сообщений в сюжетной цепочке, что отражает общий интерес к данной теме, и временем публикации сообщений.

Сюжетная цепочка выстраивается в результате отработки пользовательского запроса, процесс составления которого в этом случае максимально упрощается - для получения точных результатов вполне достаточно указать одно-два слова, относящихся к необходимой тематике.

Вместе с тем прогнозный вопрос, состоящий в том, что о событии пишут пока мало, но оно важное и в дальнейшем получит большой резонанс остается открытым. Этот вопрос связан с общей задачей нахождения исключений или аномалий, т.е. объектов, которые своими характеристиками значительно выделяются из общей массы (хотя в дальнейшем могут породить множество себе подобных).

Подход Солтона к выявлению новых событий

Подход Солтона [31], [58] заключается в использовании векторно-пространственного представления документов и традиционных методов кластеризации. При этом малый вес приписывается высокочастотным словам из массива документов, что вполне укладывается в модель TF*IDF. Напомним, TF – это локальная частота терма (Term Frequency), а IDF – величина, обратная частоте встречаемости во всем потоке документов, содержащих данный терм (Inverse Document Frequency).

В то время как локальная частота терма в документе говорит о значимости терма в пределах документа, то обратная частота встречаемости свидетельствует об уникальности терма во всем потоке документов. Поэтому произведение этих величин – достаточно удачный критерий определения веса терма.

Документы при этом подходе обрабатываются последовательно в соответствии с таким алгоритмом:

1. Первому рассматриваемому документу ставится в соответствие первый кластер. Каждый кластер представляется вектором термов (ключевых слов), входящих в документы этого кластера. Нормированный каким-то образом вектор термов принято называть центроидом. Иногда центроидом называют документ, самый близкий по некоторому критерию к вектору термов данного кластера, что не меняет сути данного алгоритма.
2. Каждый следующий документ сравнивается с центроидами существующих кластеров (для этого вводится некоторая мера близости).
3. Если документ достаточно близок к некоторому кластеру, то он приписывается этому кластеру, после чего происходит пересчет соответствующего центроида.
4. Если документ не сблизок к существующим кластерам, то происходит формирование нового кластера, которому приписывается данный документ.
5. Временной диапазон рассматриваемых документов принято называть «окном наблюдения». Кластеры, все документы которых выходят за пределы окна наблюдения, выносятся за рамки рассмотрения.

В результате работы алгоритма каждому новому возникающему кластеру соответствует новое событие, отражаемое в документах данного кластера.

Подход Папка (запросы)

В сответствии с подходом, предлагаемым Р. Папком [61], новые события выявляются из документов, не удовлетворяющих запросам пользователей, построенным с учетом уже известных событий. Алгоритм выявления новых событий заключается в следующем:

1. Формируются запросы по известным темам (при этом используются техники Text Mining – выявления и выбора понятий из текстов сообщений).

2. Новый поступающий документ сравнивается с существующими запросами.
3. Если документ не соответствует запросам, то он ассоциируется с новым событием.
4. В систему включается новый запрос, соответствующий данному документу (опционально).

Дополнительно к приведенному алгоритму подход Папка подразумевает использование механизмов ранжирования результатов поиска (для выбора центроидов кластеров), выявления и выбора понятий [40], а также архитектуру системы избирательного распространения информации InRoute [47].

Многопараметрический подход в рамках системы InfoStream

Предлагаемый авторами подход базируется на таких предположениях, относящихся к публикации информации о новых событиях:

а) минимальное время, прошедшее с момента публикации (это предположение базируется на последовательном рассмотрении входящих в систему новостных документов, а также на анализе дат, указанных в текстах самих документов);

б) минимизации веса термов, входящих в документ, по частотному словарю, сформированному на основании анализа массива документов в рамках окна наблюдения (это условие, аналогичное максимизации параметра IDF в векторно-пространственной модели);

в) максимизации суммарного веса термов, входящих в документ, по плюс-словарю (содержащему важные для содержания новостей слова типа «теракт», «конфликт», «сенсация» и т.п.);

г) ранг «авторитетности» источника (как правило, определяемый экспертами).

Если ввести обозначения:

n – величина окна наблюдения потока новостей;

D_1 – текущий документ;

D_n – последний документ из окна наблюдения;

D_i – i-й документ;

PlusDic – плюс словарь;

$sim(D_i, D_j)$ – мера близости документа i документу j;

$sim(D_i, PlusDic)$ – мера близости документа i плюс словарю,

то второе и третье условия можно записать следующим образом:

$$sim(D_i, PlusDic) > \alpha,$$

$$\sum_{j=2}^{n} sim(D_i, PlusDic) > \beta,$$

где α и β – эмпирически определяемые параметры. При этом, если рассматривать только высокоранговые источники, то эти два условия на практике оказываются вполне достаточными для выявления новых событий.

Мера близости $\mathrm{sim}(D_i, D_j)$ может быть определена традиционно для векторно-пространственой модели. Пусть $D_i = \{w_{ik}\} = \{ w: w \in D_i\}$ - документ, рассматриваемый как множество термов “Bag of Words”, $D_i + D_j = \{w: w \in D_i \mid w \in D_j\}$ – объединение термов из документов D_i и D_j – вектор размерности N.

Определим вектор $E_i = \{e_{ik}\}$ размерности N, соответствующий документу D_i, следующим образом:

$e_{ik} = 1$, если $w_{ik} \in D_i$

$e_{ik} = 0$, иначе.

В этом случае мера близости задается формулой:

$$sim(D_i, D_j) = \frac{\sum_{k=1}^{N} e_{ik} e_{jk}}{N}.$$

В [70] дано еще одно определение меры близости документов, использующее аппарат условных вероятностей, а именно, вероятность того, что некоторое слово w входит в документ D_i при условии, что оно входит в документ D_j:

$sim(D_i, D_j) = \mathrm{Prob}(w \in D_i \mid w \in D_j)$.

Очевидно, что при подходе к документу как к множеству независимо входящих в него термов обе приведенные выше формулы эквивалентны.

Предлагаемый авторами общий критерий выявления новых событий, учитывающий условия а) - г) может быть записан следующим образом:

$$\frac{Rang_i * sim(D_i, PlusDic)}{\log(i+1)\sum_{j=1}^{N} sim(D_i, D_j)},$$

где $Rang_i$ – ранг источника, соответствующего i-му документу.

Задачи выявления, отслеживания и группировки событий на основе анализа новостей активно обсуждаются специалистами во всем мире в течение уже нескольких лет. Как выяснилось, они имеют большое практическое значение имеенно сегодня, когда режим онлайн-доступа к системам интеграции новостей существенно облегчен.

Опыт разработки и внедрения элементов определения и группировки новых событий в рамках технологии показал показал свою эффективность как существенное дополнение к поисковым режимам. При этом самое важное, пожалуй то, что пользователь привязывается не к новым документам, а к новым событиям реального мира, т.е. происходит своеобразный «семантический сдвиг» восприятия. Пользователь буквально привязывается к экрану дисплея, время от времени нажимая на клавишу ”New Event”, постоянно обновляя список новостей, отслеживая в режиме реального времени важнейшие события.

8. Проблема выявления тональности сообщений

Темпы развития, динамика и объемы информационного пространства сети Интернет превращают его в информационный поток [1]. Исследование потока новостных сообщений, публикуемых на страницах Web-сайтов, должно использовать принципиально новый инструментарий, так как традиционные методы не всегда способны охватить даже репрезентативную часть этого потока (о полноте в данном случае вообще не может быть и речи). Традиционная эксперная оценка текстовых сообщений оказывается не эффективной для сверхбольших и сверхдинамичных текстовых массивов. В рамках этой монографии рассматривается один из аспектов анализа текстов сообщений из современных информационных потоков, а именно, оценка тональности текстов. Под тональностью текста в контексте данной работы понимается позитивная, негативная или нейтральная эмоциональная окраска как всего сообщения, так и отдельных его частей, имеющих отношения к определенным понятиям, таким как персоны, организации, бренды и т.п.

Самая известная разработка в области автоматизации процесса определения тональности текстов – это российская система «ВААЛ» (http://www.vaal.ru/) [8], ориентированная на эмоционально-лексический анализ. Эта система базируется на построении частотного словаря, его анализа на присутствие определенных слов, позволяющих определить с некоторой вероятностью психолингвистические категории.

Другой подход, ориентированный на использование лингвистических алгоритмов, статистического разбора текстов, учитывающий модальные характеристики ситуации, модусные смыслы и отношение автора к описываемой ситуации [11], является перспективным, но достаточно ресурсозатратным и далеко не универсальным.

Вместе с тем характер большинства публикаций электронных СМИ дает возможность оценивать тональность текста, его эмоциональную окраску

непосредственно по составу лексики, что близко к подходам, применяемым в системе «ВААЛ».

Метод, реализация которого представлена в данной работе, основывается на статистических закономерностях, связанных с присутствием определенных лексем в текстах, наивном байесовском подходе и методе нейронных сетей (реализации двухслойного перцептрона). Отличительной чертой данного метода является его простота и универсальность, при том, что точность оценки может регулироваться параметрически в достаточно широком диапазоне. Близкие по идеологии подходы сегодня широко применяются для борьбы со спамом и дают впечатляющие результаты [52], [53].

Однако необходимо заметить, что задача определения тональности сообщений более сложна, чем выявление спама на основе анализа текстов. В то время как выявление спама подразумевает лишь две гипотезы (спам, не спам), то в задаче определения тональности проверяется как минимум три: эмоциональная окраска позитивная, негативная, нейтральная и, зачастую, существует потребность также в проверке комбинации этих гипотез (например, для выявления уровня «экспрессивности» текста).

С другой стороны, в отличие от проблемы выявления спама, где оценка отдельных документов может быть близка к однозначной, в случае определения тональности сообщений даже разные люди – эксперты не всегда приходят к единому мнению. Тут ситуация близка к оценке уровня релевантности-пертинентности при информационном поиске [30], [18].

Поскольку предложенный подход близок к тому, который используется в байесовских антиспамовских фильтрах, остановимся в начале на применении теоремы Байеса к решению проблемы спама. В методе Байеса подразумевается использование оценочной базы — двух корпусов электронных писем, один из которых составлен из спама, а другой — из обычных писем. Для каждого из 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 (Paul Graham), если сообщение содержит n слов с весовыми оценками $w_1...w_n$, то оценка условной вероятности того, что письмо окажется спамом, основанная на данных из оценочных корпусов, вычисляется по формуле:

$$Spm = \Pi w_i / (\Pi w_i + \Pi(1-w_i)). \qquad (10.1)$$

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:

$$P(S \mid A) = \frac{P(A|S)P(S)}{P(A|S)P(S)+P(A|\bar{S})P(\bar{S})} \quad . \qquad (10.2)$$

Если изначально не известно, является письмо спамом или нет, исходя из опыта предполагается, что $P(\bar{S}) = \lambda P(S)$, на основании чего из (10.2) следует:

$$P(S \mid A) = \frac{P(A|S)}{P(A|S)+\lambda P(A|\bar{S})} \quad . \qquad (10.3)$$

Далее формула (10.3) 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:

$$P(S \mid A_1 \,\&\, A_2) = \frac{P(A_1|S)P(A_2|S)}{P(A_1|S)P(A_2|S)+\lambda P(A_1|\bar{S})P(A_2|\bar{S})} = \frac{p(t_1)p(t_2)}{p(t_1)p(t_2)+\lambda(1-p(t_1))(1-p(t_2))} \quad . \qquad (10.4)$$

Обобщением формулы (10.4) на случай произвольного количества слов и λ=1 является формула (10.1). Следует заметить, что широко применяемое в антиспамовских фильтрах находит именно значение λ=1. С одной стороны, это нексколько упрощает вычисления, но с другой – серьезно искажает действительность и существенно снижает качество работы этих программ.

На практике на основе словарей, которые постоянно модифицируются, для каждого сообщения рассчитывается значение *Spm*. Если оно больше некоторого порогового, то сообщение считается спамом.

В случае оценки тональности сообщений пространство гипотез будет содержать: H_{-1} – тональность отрицательная, H_0 - тональность нейтральная и H_1 - тональность положительная. Для упрощения рассмотрим события такого типа: H_1 – тональность положительная, $\overline{H_1}$ – тональность не положительная. Из корпуса документов с положительной тональностью выбираются слова, характерные для этих документов. Из них выбираются слова t со значениями $p(t|H_1)$, превышающими ½, например 0,6. Такие слова принято называть тонально-окрашенными или просто тональными, несущими в себе оценочную семантику.

Для упрощения модели предположим, что для всех выбранных термов вес будет одинаковым, равным α (может изменяться при обучении модели). Тогда формула (10.1) примет вид:

$$Spm(x) = \frac{\alpha^x}{\alpha^x + \lambda(1-\alpha)^x}, \qquad (10.5)$$

где x – число весомых с точки зрения тональности слов в информационном сообщении, α – вес.

Как видно из рис. 44 (α = 0,6, λ = 1), наличие 10 слов, характерных для положительной тональности, практически гарантирует, что все сообщение будет обладать этим же свойством.

Для оценки гипотезы об отрицательной тональности сообщения (H_{-1}) может использоваться словарь слов «отрицательной тональности» и та же формула (10.5). Вместе с тем поскольку положительная и отрицательная тональности являются своего рода антагонизмами, окончательное решение о тональности сообщения принимается с учетом разности значений весовых оценок гипотез H_1 и H_{-1}. Пороговое значение этой величины - β определяется в процессе настройки (обучения) системы.

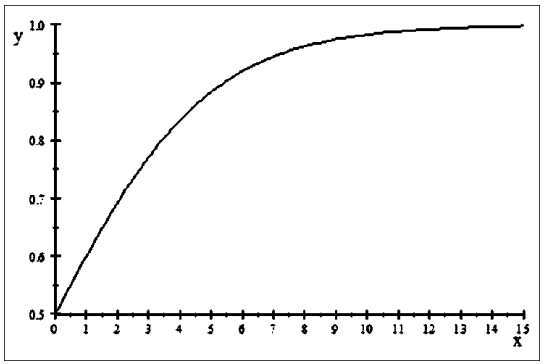


Рис. 44. График функции Spm(x)

Необходимо сделать еще одно, диктуемое практикой, замечание. Следует учитывать, что отрицательная тональность сообщений в Интернет почти всегда выражена более явно, чем положительная. Поэтому для сопоставимости тональностей при расчете веса отрицательной тональности значение x в формуле (10.5) несколько уменьшается путем умножения его на эмпирически определяемую константу $\in (0,1)$.

В некоторых случаях определенный интерес для аналитиков представляют документы, у которых достаточно высоки значения весов как положительной, так и отрицательной тональности. Заметим, что разница этих весов может быть минимальной, т.е. документ может даже характеризоваться как нейтрально окрашенный. Вместе с тем он получит характеристику «экспрессивной» тональности.

Реализацию приведенного алгоритма представим в виде нейронной сети [33] (рис. 45). Первый слой этой сети составляют два нейрона – определители весовых значений положительной и отрицательной тональности (положительный и отрицательный нейроны). Можно предположить, что количество дендритов каждого нейрона равно количеству слов из словаря естественного языка. На вход нейронов поступают входные сигналы - значения $x_1 \ldots x_n$, соответствующие входным словам. При этом x_i=1, если на вход поступило слово из словаря с номером i, в противном случае x_i=0. Весовые значения (веса синапсов), которые

соответствуют этим словам, равны $w^+_1 \ldots w^+_n$ для положительного нейрона и $w^-_1 \ldots w^-_n$ - для отрицательного. Именно эти весовые значения могут изменяться в процессе обучения перцептрона. Сумматоры подсчитывают значения NET^+ и NET^-, соответственно. Проводимость нейронов рассчитывается по формуле (10.5), аргументом в которой выступает значение NET^+ для положительного нейрона и γ NET^- – для отрицательного. Оба нейрона выдают через аксоны градиентные значения, OUT^+ и OUT^-, которые являются входными сигналами для нейрона второго уровня, сумматор которого вычисляет разность OUT^+ и OUT^-, а функция проводимости выдает градиентный результат по условию, приведенному на рис. 45.

Представленная модель реализована в системе контент-мониторинга InfoStream, которая применяется для решения задач автоматизированного сбора новостной информации с открытых Web-сайтов, ее систематизации и обеспечения доступа к ней в поисковых режимах. Эта система в настоящее время охватывает свыше 2000 источников – более 40000 уникальных новостных сообщений в сутки, при этом в ее ретроспективных базах данных хранится свыше 30 млн. записей. Для навигации в этих информационных ресурсах и для уточнения запросов был разработан механизм информационного портрета, который представляет собой многоаспектную подборку параметров выборки по первоначально составленному запросу [4]. В информационном портрете в качестве одного из параметров используется значение тональности сообщения, уточнение по которому позволяет выделить публикации негативной или позитивной тональности, соответствующие тематике, определенной первоначально введенным запросом.

Еще одна возможность системы InfoStream – отслеживание динамики появлений понятий, соответствующих введенным пользователем запросам. На рис. 46. представлена соответствующая диаграмма появления сообщений по запросу “Microsoft” в течение семи дней. Сообщения, отмеченные позитивной или негативной тональностью, образуют на представленной диаграмме отрезки разных цветов. Данная выходная форма системы контент-мониторинга InfoStream пользуется большой популярностью в аналитических службах.

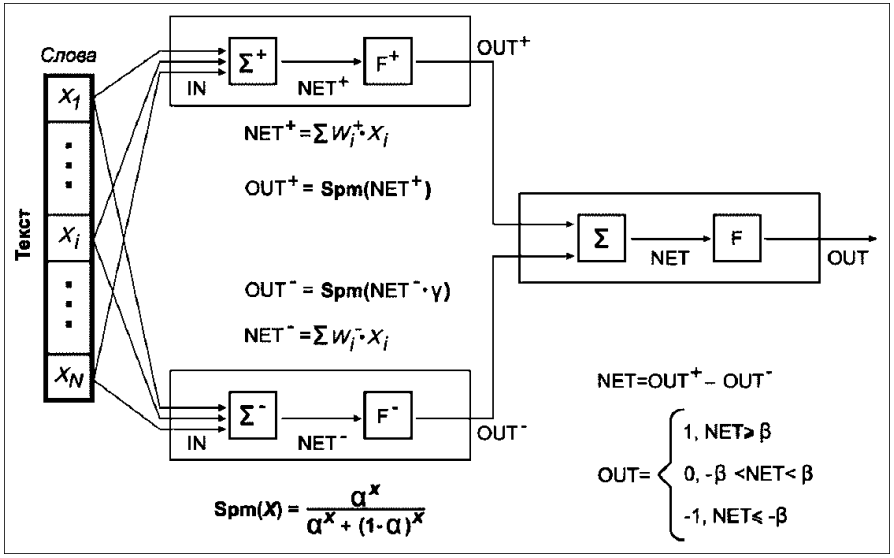


Рис. 45. Двухслойный перцептрон определения тональности текста

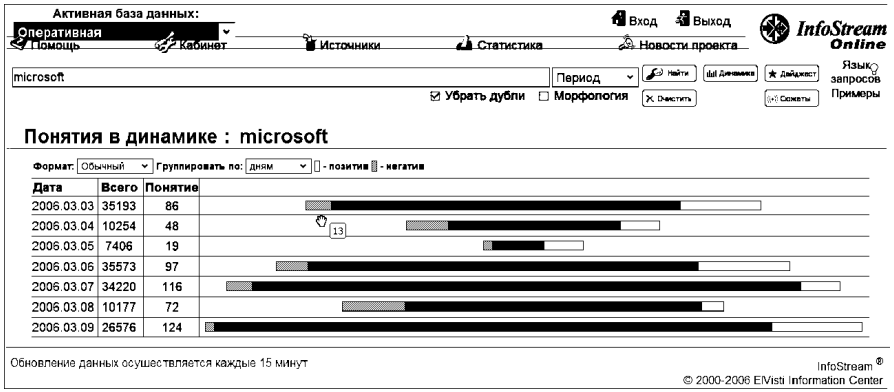


Рис. 46. Динамика появления понятий с указанием количества соответствующих им сообщений с позитивной и негативной тональностью

Предложенный алгоритм используется как инструмент для определения общей тональности сообщений, в то время как тональность отдельных понятий, охватываемых сообщением, не всегда соответствует тональности всего сообщения. Можно предположить (и такое предположение, по мнению

экспертов-аналитиков, вполне оправдывается на практике), что при анализе достаточно большого информационного канала, соответствующего некоторому запросу, эмоциональная окраска целевого понятия будет вполне соответствовать интегральной оценке всего информационного канала.

Более точные результаты могут быть получены путем оценки не всего сообщения, а его фрагмента, например абзаца, включающего интересующее понятие, предложения или даже части предложения.

Конечно, такой подход не гарантирует точности в каждом конкретном случае, а дробление текста сообщения может привести и к потере его полноты. Однако, повторим, для репрезентативных информационных каналов, относящихся к целевому понятию, представленная методика, в силу статистических закономерностей, оказывается не только «прозрачной», но и достаточно эффективной.

Заключение

Системы синдикации Интернет-новостей решают проблему нахождения необходимой информации, но оставляют без внимания такие задачи, как обобщение данных - их обработку и анализ. Одним из самых перспективных направлений обобщения информационных потоков в настоящее время является метод «глубинного анализа текстов» (Text Mining). Применительно к новостным потокам его идеологию можно сформулировать как постоянное воспроизводимое во времени выполнение их содержательного анализа. Именно непрерывная аналитическая обработка сообщений является самой характерной чертой этого метода, который позволяет формировать автоматические дайджесты, выявлять новые понятия и их взаимосвязи, рассчитывать разнообразные рейтинги. Именно системы такого типа смогут избавить пользователей от дублирующейся информации, информационного шума, позволят выявлять главные тенденции, находить коррелирующие сообытия, аномалии.

В настоящее время выделяют четыре основных вида приложений технологий Text Mining:

- Классификация текстов благодаря выявлению статистических корреляций для формулирования правила размещения документов в предопределенные категории.
- Кластеризация, базирующаяся на выявлении латентных признаков документов, использующая лингвистические и математические методы без использования предопределенных категорий, что может дать эффективный охват больших объемов данных.
- Построение семантических сетей на основе анализа документальных информационных потоков.
- Извлечение фактов из текстов.

Уже из одного перечня названий видно, что к собственно знаниям эти приложения имеют весьма отдаленное отношение. Их, скорее, можно рассматривать как некую промежуточную платформу, облегчающую дальнейшие

манипуляции с данными. Например, в плане поставленной цели извлечение фактов из текста имеет смысл лишь в том случае, когда предвидится дальнейшее установление определенных отношений между ними – разрозненные факты, лишенные связей, претендовать на знания ни в коей мере не могут.

Можно назвать еще несколько задач технологии Text Mining, например, прогнозирование, которое состоит в том, чтобы предсказать по значениям одних признаков объекта значения остальных.

Еще одна задача — нахождение исключений или аномалий, т.е. поиск объектов, которые своими характеристиками сильно выделяются из общей массы. Для этого сначала выясняются средние параметры объектов, а потом исследуются те объекты, параметры которых наиболее сильно отличаются от средних значений. Сегодня подобный анализ часто проводится после классификации, для того чтобы выяснить, насколько последняя была точна.

Здесь мы уже видим отчетливые, хотя и явно недостаточные элементы перехода от формальных систем к содержательным.

Несравненно ближе к решению общей проблемы извлечения знаний из текста стоит задача поиска скрытых связей отдельных признаков (дескрипторов, понятий). От предсказания эта задача отличается тем, что заранее неизвестно, по каким именно признакам реализуется взаимосвязь; цель именно в том и состоит, чтобы найти связи признаков. Решение этой задачи позволяет сокращать размерность пространства признаков, создавать обозримые классификаторы, пригодные для решения задач навигации в информационных потоках.

Одна из задач, которая решается в рамках концепции Text Mining – это автоматическое реферирование. В случае информационных потоков – составление дайджестов на основании нескольких документов. Можно ли построить реферат документа, не зная языка, на котором он составлен? Оказывается, можно, при условии, что он удовлетворяет законам Зипфа!

Это кажется невероятным. Вместе с тем мы можем передать основное содержание документа, не имея ни малейшего представления о том, что в нем написано, и мы это действительно делаем. То есть, машина не просто

экстрагирует фрагменты из документа, но эти фрагменты на самом деле содержат в себе самое важное, что есть в документе.

При этом машина «не знает» не только значения слов, которыми она манипулирует, но даже того, что последовательности символов, разделенные пробелами на самом деле являются словами. Все, что ей нужно, это определенные закономерности, существующие в тексте между повторяющимися последовательностями символов. Машина, пользуясь некоторым лингвостатистическим алгоритмом, выбирает заданное число таких последовательностей и выдает их в определенном порядке. И оказывается, что человек, владеющий соответствующим языком, согласится, что представленные машиной данные дейсвительно передают содержание документа. (Мы хорошо знаем, как это происходит, но не имеем ни малейшего понятия – почему.)

И наконец, для обработки и интерпретации результатов Text Mining большое значение имеет визуализация. Визуализация на основе систем Text Mining предполагается как средство представления контента всего потока документов, а также для реализации навигационного механизма, который может применяться при исследовании документов и их классов.

Рискнем предположить, что это – наиболее выдающееся достижение современных информационных технологий в данном направлении. Суть заключается в том, что эффективное представление потоковых данных в удобной для пользователя форме позволяет непосредственно задействовать человеческий интеллект, который, в конечном счете, намного быстрее приведет к поставленной цели, чем любая машина. Собственно, машина и должна предоставить пользователю в удобной форме то, с чем он в общем случае сможет справиться и сам. Как бы там ни было, приходится признать, что проблема извлечения знаний из текста находится, видимо, на стадии осмысления и, вероятно, в ближайшее время начнет бурно развиваться.

Одна из актуальнейших задач, стоящих перед учеными различных специальностей, состоит в построении четкой модели современного информационного пространства, которая базируется на достижениях в области

лингвистики и информатики, а также на методах, близких к методам теоретической физики, строгом математическом инструментарии. В частности, предполагается дальнейшее развитие обучаемых алгоритмов, которые в противоположность традиционным концепциям искусственного интеллекта должны обеспечить возможность построения рекуррентных процедур с интерактивным участием человеческого интеллекта.

Вместе с тем исследования современных информационных потоков могут представлять немалый интерес как для лингвистов, математиков, так даже и для физиков, например, в плане аналогового моделирования статистических процессов, в том числе сложных нелинейных систем с элементами самоорганизации. Семантика информационного пространства обуславливает и развитие новых методов кодирования и сжатия информации, включая средства обеспечения однозначности дешифровки сообщений.

Предполагается, что новая ступень развития Web-пространства будет определяться технологиями работы с огромным объемом информации, накопившимся в Интернет. Web следующего поколения будет характеризоваться переходом от сети документов к сети данных, которые при необходимости агрегируются в семантически связанные документы с помощью Web-сервисов. Предполагается также существование единого информационного пространства в виде множества единиц данных, которые могут размещаться на многочисленных сайтах в Интернете. Пользователь будет получать документ путем агрегирования у себя на рабочем месте этих информационных единиц.

Перспективы охвата информационного пространства, по-видимому, будут зависеть от создания и развития эффективной инфраструктуры, в рамках которой будут работать программные продукты со стороны Web-серверов и пользователей.

Даже частичное решение названных задач при наличии обширной и дешевой экспериментальной базы позволит уже в настоящее время реализовать полезные и эффективные инструменты работы и серфинга в информационных потоках.

Литература

[1] Брайчевский С.М., Ландэ Д.В. Современные информационные потоки: актуальная проблематика. // Научно-техническая информация. - Сер. 1. - 2005. - №11, - С. 21-33.

[2] Вольтерра В. Математическая теория борьбы за существование. М.: Наука, 1976.

[3] Горькова В. И. Информетрия (Количественные методы в научно-технической информации). // Итоги науки и техники. Сер. Информатика. - Т.10. – М.: ВИНИТИ, 1988.

[4] Григорьев А.Н., Ландэ Д.В. Адаптивный интерфейс уточнения запросов к системе контент-мониторинга InfoStream. // Труды Международного семинара «Диалог'2005». – 2005. – С. 109-111.

[5] Григорьев А.Н., Ландэ Д.В. Система мониторинга новостей InfoStream - информационное пространство из одних рук. Построение информационного общества: ресурсы и технологии. // Тезисы докладов и информационные материалы XI международной научно-практической конференции. - К:УкрИНТЭИ. - 2005. - С. 17-20.

[6] Грушо А.А., Тимонина Е.Е. Теоретические основы защиты информации - М.: Агентство "Яхтсмен", 1996.

[7] Додонов А.Г., Клещев Н.Т., Клименко В.Г. Анализ отраслевых вычислительных сетей. - Л.: Судостроение, 1990. - 256 с.

[8] Дымшиц М. Репрезентационные системы. // Руководство к использованию программы ВААЛ. − М., 1999.

[9] Дюбуа П. MySQL. - М.: ИД Вильямс, 2004. - 1056 с.

[10] Ефимов А.Н. Информация: ценность, старение, рассеяние. - М., 1978.

[11] Ермаков А.Е., Киселев С.Л. Лингвистическая модель для компьютерного анализа тональности публикаций СМИ. // Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции Диалог'2005. – Москва, Наука, 2005. – С. 282-285.

[12] Иванов С.А. Стохастические фракталы в информатике. // Научно-техническая информация. - Сер. 2. – 2002. - № 8. – С. 7-18.

[13] Иванов С.А., Круковская Н.В. Статистический анализ документальных информационных потоков. // Научно-техническая информация. Информ. процессы и системы. — Сер. 2. — 2004. — № 2. — С. 11–14.

[14] Ландэ Д.В. Данные в кармане. // ЧИП-Украина. – 2002. - № 6. – С. 82 – 85.

[15] Ландэ Д.В. Затерянный вэб. // Телеком. - 2005. - №. 1. - С. 46-51.

[16] Ландэ Д.В. На границе стихий. // ЧИП-Украина. - 2003. - №. 5. - С. 72-77.

[17] Ландэ Д.В. Некоторые методы анализа новостных информационных потоков. // Научные труды Донецкого национального технического университета. Серия: Информатика, кибернетика и вычислительная техника (ИКВТ-2005). - Вып. 93. – Донецк: ДонНТУ, 2005. - С. 277-287.

[18] Ландэ Д.В. Поиск знаний в Internet. Профессиональная работа - М.: ИД “Вильямс”, 2005. - 272 с.

[19] Ландэ Д.В. Семантический вэб: воплощение идеи. // "Телеком". - 2005. -№ 6. - С. 60-65.

[20] Ландэ Д.В. Сканер системы контент-мониторинга InfoStream. // Открытые информационные и компьютерные интегрированные технологии: Сб. науч. трудов. – Харьков: НАКУ «ХАИ», 2005. – Вып. 28 - С. 53-58.

[21] Ландэ Д.В., Брайчевский С.М. Определение тематической направленнос-ти запросов путем анализа набора рейтинговых источников. // Открытые информационные и компьютерные интегрированные технологи: Сб. научн. трудов. Вып. 29. – Харьков: Нац. аэрокосмический ун-т «Хаи», 2005. - С. 169-174.

[22] Ландэ Д.В., Брайчевский С.М., Григорьев А.Н. Прогнозно-аналитические исследования на основе системы контент-мониторинга InfoStream. // Тезисы докладов V международной научно-практической конференции «Информация, анализ, прогноз - стратегические рычаги эффективного государственного управления». -К:УкрИНТЭИ, 2006. - С. 147-152.

[23] Ландэ Д.В., Литвин А.Б. Феномены современных информационных потоков. // "Сети и бизнес". - 2001. - № 1. - С. 14-21.

[24] Ландэ Д.В, Морозов А.Ю. Новостной Интернет. // Телеком, -№ 1-2. – 2005. - С. 58-62.

[25] Ландэ Д.В, Морозов А.Ю. Читайте новости, батенька! // ЧИП-Украина. – 2004. - №. 7. - С. 82-85.

[26] Ландэ Д.В., Фурашев В.Н. Вопросы построения и использования многокритериальной модели выбора источников информации. // Открытые информационные и компьютерные интегрированные технологии: Сб. науч. трудов. Вып. 30. –Х.: аэрокосмический ун-т "ХАИ", 2006. - С. 76-85.

[27] Ланде Д.В., Фурашев В.М, Григор’єв О.М. Програмно-апаратний комплекс інформаційної підтримки прийняття рішень: Науково-методичний посібник. - К.: ТОВ "Інжиніринг", 2006. - 48 с.

[28] Мотылев В. М. Старение научно-технической литературы. – Л., Наука, 1986.

[29] К. Нейл, Г. Шанмагантан. Web-инструмент для выявления плагиата. // - Открытые системы. -2005. -№ 1.

[30] Сегалович И.В. Как работают поисковые системы. // Мир Internet. – 2002. -№ 10.

[31] Солтон Дж. Динамические библиотечно-информационные системы. – М.: Мир, 1979. - 560 с.

[32] Уильямсон М. Анализ биологических популяций. - М.: Мир, 1975.

[33] Уссермен Ф. Нейрокомпьютерная техника. / - М.: «Мир», - 1992. – 184 с.

[34] Уэйнрайт П. Apache для профессионалов. –М.: Лори, Wrox Press Ltd, 2001. – 474 с.

[35] Федер Е. Фракталы. —М.: Мир, 1991. — 254 с.

[36] Фурашев В.М., Ланде Д.В., Григор'єв О.М., Фурашев О.В. Електронне інформаційне суспільство України: погляд у сьогодення і майбутнє. / Академія правових наук України. Науково-дослідний центр правової інформатики. - К.: Інжиніринг, 2005. — 163 с.

[37] Фурашев В.М., Ланде Д.В., Брайчевский С.М. Системная информатизация избирательных и референдумных процессов: методологические основы статистических исследований электронных информационных ресурсов в период избирательной кампании. // Открытые информационные и компьютерные интегрированные технологии: Сб. науч. трудов. Вып. 29. –Х.: аэрокосмический ун-т "ХАИ", 2005. - С. 11-15.

[38] Шапошников И. Web-сервисы Microsoft .NET. // - СПб.: БХВ-Петербург". - 2002. - 334 с.

[39] Шрейдер Ю.А. Равенство, сходство, порядок. - М.: "Наука", 1971. - 256 с.

[40] Allan J. Incremental Relevance Feedback for Information Filtering, In Proccedings of ACM SIGIR, 270-278, 1996.

[41] Baeza-Yates R. and Ribeiro-Neto B. (1999), Modern Information Retrieval, Addison-Wesley, 1999.

[42] Tim Berners-Lee, James Hendler, Ora Lassila, The Semantic Web, Scientific American, May 2001 (http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21).

[43] Andrei Z. Broder. Identifying and Filtering Near-Duplicate Documents, COM'00 // Proceedings of the 11th Annual Symposium on Combinatorial Pattern Matching. – 2000. p 1-10.

[44] Andrei Z. Broder, Steven C. Glassman, Mark S. Manasse. Syntactic Clustering of the Web // WWW6, 1997.

[45] Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, Janet Wiener. *Graph structure in the web* (http://www.almaden.ibm.com/cs/k53/www9.final/).

[46] Burton R.E. and Kebler R.W. The "half-life" of some scientific and technical literatures. American Documentation 1960;1:98—109.

[47] Callan, J. Document Itering with Inference Networks, In Proccedings of the 19th Annual International ACM SIGIR Cjnference on Research and Development in Information Retrieval, 262-269, 1996.

[48] Cole P.F. Journal usage versus age of journal // J.Doc. – 1963. – Vol. 19, №1. – P. 1-10.

[49] G. M. Del Corso, A. Gulli, and F. Romani. Ranking a stream of news. In Proceedings of 14th International World Wide Web Conference, pages 97–106, Chiba, Japan, 2005.

[50] Gianna M. Del Corso, Antonio Gulli, Francesco Romani: Fast PageRank Computation Via a Sparse Linear System (Extended Abstract). WAW 2004: 118-130.

[51] Department of Defense Trusted Computer System Evaluation Criteria - DoD, 1985.

[52] P. Graham, Better Bayesian Filtering. http://paulgraham.com/better.html, January 2003.

[53] P. Graham, A Plan for Spam. http://paulgraham.com/spam.html, August 2002.

[54] A Guide to Understanding Covert Channel Analysis of Trusted Systems, NCSC-TG-030, ver. 1 - National Computer Security Center, 1993.

[55] Handbook for the Computer Security Certification of Trusted Systems - NRL Technical Memorandum 5540:062A, 12 Feb. 1996.

[56] H.S. Heaps. Information Retrieval: Computation and Theoretical Aspects, pages 206-208. Academic Press, Inc., Orlando, FL, 1978.

[57] S. Ilyinsky, M. Kuzmin, A. Melkov, I. Segalovich. An efficient method to detect duplicates of Web documents with the use of inverted index // WWW2002, 2002.

[58] Kurt, H. On-line New Event Detection and Tracking in A Multi-Resource Environment, MS. Thesis, Bilkent University, 2001.

[59] Malthus T.R. An essay on the principal of Population . 1798 (Penguin Books 1970).

[60] U. Manber. Finding similar files in a large file system. Proceedings of the 1994 USENIX Conference, pp. 1-10, January 1994.

[61] Papka, R. On-line News Event Detection, Clustering, and Tracking. Ph. D. Thesis, University of Massachusetts at Amherst, September 1999.

[62] Pearl R. The Introduction to Medical Biometry and Statistics. Philadelpia, 1930; Ibid. The Natural History of Population. L., 1939.

[63] Chris Preimesberger. Web 2.0: Possibly the best IT business conference of 2004 // NewsForge, 2004.

[64] Van Raan A.F.J. Fractal Geometry of Information Space as Represented by Cocitation Clustering // Scientometrics. —1991. — Vol. 20, N 3. — P. 439–449.

[65] G. Salton, J. Allan, and C. Buckley. Approaches to passage retrieval in full text information systems. In ACM SIGIR conference on R&D in Information Retrieval, pages 49-58, 1993.

[66] Chakrabarti Soumen, Mining the web. Discovery knowledge from hypertext data// Publisher: Morgan Kaufmann, 2002. 344 p.

[67] W.R. Stone. Plagiarism, Duplicate Publication and Duplicate Submission: They Are All Wrong! // IEEE Antennas and Propagation, Aug. 2003. -Vol. 45. -№ 4.

[68] Danny Sullivan. Invisible Web Gets Deeper. // The Search Engine Report. – 2002. (http://searchenginewatch.com/sereport/article.php/2162871).

[69] Verhulst P.F. Notice sur la loi que la population suit dans son accroissement Corr. Math. Et Phys. 10, 113-121, 18.

[70] Vutal, A. Online New Event Detection and Clustering using the Concepts of the Cover Coefficient-based Clustering Methdology. MS Thesis, Bilkent University, 2002.

[71] Zipf, George Kingsley (1949): Human behavior and the principle of least effort. Wesley, Cambridge, MA, 1949.