А.Г. ДОДОНОВ, Д.В. ЛАНДЭ,
В.В. ПРИЩЕПА, В.Г. ПУТЯТИН

КОМПЬЮТЕРНАЯ КОНКУРЕНТНАЯ РАЗВЕДКА

НАЦИОНАЛЬНАЯ АКАДЕМИЯ НАУК УКРАИНЫ
ИНСТИТУТ ПРОБЛЕМ РЕГИСТРАЦИИ ИНФОРМАЦИИ

А.Г. Додонов, Д.В. Ландэ,

В.В. Прищепа, В.Г. Путятин

КОМПЬЮТЕРНАЯ КОНКУРЕНТНАЯ РАЗВЕДКА

Киев – 2021

УДК 004.5
ББК 22.18, 32.81, 60.54
С95

А.Г. Додонов, Д.В. Ландэ, В.В. Прищепа, В.Г. Путятин
Компьютерная конкурентная разведка. – Киев: ТОВ «Інжиніринг», 2021. – 354 с.

Книга посвящена рассмотрению вопросов компьютерной конкурентной разведки, разведки в открытых ресурсах сети Интернет. Компьютерная конкурентная разведка охватывает автоматизированные процедуры сбора и аналитической обработки информации, которые проводятся с целью поддержки принятия управленческих решений, повышения конкурентоспособности исключительно из открытых источников в компьютерных сетях – веб-сайтов, блогосферы, социальных сетей, мессенджеров, баз данных. В книге рассматриваются различные вопросы информационно-аналитической деятельности в сетевой среде. В качестве теоретических основ компьютерной конкурентной разведки рассматриваются элементы теории информации, анализа социальных сетей, информационного и математического моделирования.

Для широкого круга специалистов в области информационных технологий и безопасности.

Рекомендовано к изданию ученым советом Института проблем регистрации информации НАН Украины (протокол № 9 от 17 февраля 2021 года)

Рецензенты:
Член-корр. НАН Украины, д.т.н., профессор *В.В.Мохор*
Д.т.н., профессор *А.Я. Матов*
Д.ю.н., профессор *К.И. Беляков*

ISBN 978-966-2344-79-0

Оглавление

Введение

Компьютерная конкурентная разведка (Computer Competitive Intelligence) охватывает процедуры сбора и обработки информации, проводимые с целью поддержки принятия управленческих решений, повышения конкурентоспособности организаций исключительно из открытых источников из компьютерных сетей, большинство из которых являются надстроенными над сетью Интернет, так называемыми, оверлейными. Поэтому часто в качестве синонима конкурентной разведки используется термин интернет-разведка. Таким образом, данная книга фактически посвящена проблематике конкурентной разведки, но с одним существенным ограничением – все источники информации, необходимые для проведения разведывательной деятельности, являются открытыми и доступными в компьютерных сетях. Более того, большая часть инструментария, программ обработки информации, также свободно доступна через современные компьютерные сети. В англоязычной литературе такой вид разведки принято называть разведкой по открытым источникам (Open Sources INTelligence, OSINT) [Берд, 2007], что также можно считать синонимом термина “конкурентная разведка”. Однако следует отметить, что в зарубежной литературе употребление OSINT в значительной мере ограничено применением в государственной сфере. Но именно для технологий OSINT создано наибольшее количество методик, техник и технологий.

Разведывательная информация может быть получена из официальных источников, других открытых источников, СМИ, объявлений, рекламы, внутрифирменных, банковских, правительственных отчетов, баз данных, от экспертов, путем добывания (сбора), анализа или специальной обработки данных, текстов. Правда, при этом количество разнородных сведений, которые необходимо переработать, чтобы получить крупицы знаний огромно, а потому в настоящее время конкурентная разведка немыслима без использования специализированных информационных технологий, практического применения современной концепции больших данных (Big Data).

По мнению бывшего директора Центрального разведывательного управления США (ЦРУ) Р. Хилленкерта «80 % разве-

дывательной информации получается из таких источников как книги, журналы, научно-технические обзоры, фотографии, коммерческих аналитических отчетов, газет, теле- и радиопередач…».

По другим оценкам, в любой разведке от 35 до 95 % всей информации добывается из открытых источников. При этом доля затрат на работу с открытыми источниками, например, в разведывательном бюджете США, составляет всего лишь около 1 %.

Известно, что для бизнес-структур 95% полезной информации дает конкурентная разведка, 4,1% информации можно легально получить из государственных структур. Позволить себе полноценное проведение бизнес-разведки на рынках могут только большие компании, однако возможности конкурентной разведки доступны практически всем [https://trademaster.ua/articles/312620].

Значимость разведки по открытым источникам отметил еще президент США Линдон Джонсон (Lyndon Baines Johnson) 30 июня 1966 г., когда произнес речь на церемонии принятия присяги директором ЦРУ Ричардом М. Хелмсом (Richard McGarrah Helms): «Высшие достижения не являются результатом потихоньку пересказанной тайной информации, а следуют из терпеливого, ежечасного изучения печатных источников».

По устоявшемуся ошибочному мнению, вся полезная разведывательная информация добывается из секретных источников агентурным или оперативным путем – на самом деле это не так. Известное признание адмирала Захариаса – заместителя начальника разведки Военно-морских сил США в годы Второй мировой войны, опровергает это. Так, по его оценке 95 % информации разведка военно-морских сил черпала из открытых источников, 4 % – из официальных, и только 1 % – из конфиденциальных источников. Справедливости ради надо сказать, что часто именно этот один процент является тем золотым недостающим звеном, который позволяет сложить целостную картину разрозненной мозаики всех разведданных. И если такое соотношение справедливо для военной разведки, то тем более оно будет правильным для конкурентной разведки для бизнес-структур.

В то же время, анализ рассекреченного отчета ЦРУ за 1987 год «Enterprise-Level Computing in Soviet Economy» (SOV C87-10043) дает представление о том, какой колоссальный объем данных необходимо было обрабатывать аналитикам. Для составления отчета постоянно на протяжении года сканировалось 347 открытых источников; для создания сводки объемом в одну страницу ежедневно обрабатывался информационный массив объемом примерно 7 млн. слов.

Общеизвестно, что основное отличие конкурентной разведки от промышленного шпионажа – это легитимность и соблюдение этических норм [Дудихин, 2004]. Здесь данное положение доведено до абсолюта – исключительно все источники информации в этом случае доступны и легальны.

Интернет-разведка, разведка по интернет-источникам, как, впрочем, и вся конкурентная разведка, представляет собой особый вид информационно-аналитической работы, позволяющей собирать разностороннюю бизнес-информацию без применения тех специфических методов оперативно-розыскной деятельности, которые являются исключительной прерогативой правоохранительных органов.

Вместе с тем, методы ведения интернет-разведки, техники и технологии ее проведения весьма близки к используемым в традиционной разведывательной деятельности спецслужбами.

Применение интернет-разведки в коммерческой компании оправдывается не только соображениями информационной безопасности, но важно и для решения задач менеджмента и маркетинга тем, что обеспечивает:

- наблюдение за репутацией компании (с точки зрения клиентов, конкурентов, госорганов);
- активное участие в формировании имиджа компании, информационного поля вокруг компании;
- отслеживание появления нового конкурента, технологии или канала сбыта;
- выявление возможных слияний и поглощений;
- оценка потенциальных рисков при инвестициях;
- опережение шагов конкурентов в рамках маркетинговых кампаний;
- опережение конкурентов в тендерах;
- выявление каналов утечки информации.

Зыбкая грань между понятиями конкурентная разведка и промышленный шпионаж, состоит в легитимности, законности методов и средств, используемых в процессе сбора целевой информации [Ландэ, Прищепа, 2007]. Следует отметить также весьма тонкую разницу между бизнес-разведкой (Business Intelligence, BI) и конкурентной разведкой. Из публикаций и описаний систем, где упоминаются эти термины, можно сделать вывод, что бизнес-разведка направлена больше на изучение «внутренней» маркетинговой, финансовой, экономической информации и информации о клиентах, в то время как конкурентная разведка чаще охватывает процессы, связанные с добыванием «внешней» информации и знаний непосредственно о конкурентах.

Родоначальником современной бизнес-разведки считается компания Ксерокс (Xerox), столкнувшаяся с конкуренцией со стороны японских производителей [Прескотт, 2003]. В начале 70-х годов XX века, после выхода японцев на американский рынок, менеджеры Ксерокс заметили, что компания стала утрачивать позиции на рынке. Ситуацию исправили изменения, основанные на сборе актуальной информации о рынке и конкурентах. Ксерокс, благодаря своему японскому филиалу, создал систему оценки и анализа работы (бенчмаркинг), а затем адаптировал и применил к бизнесу разведывательные технологии. При этом одним из основных условий организации этого процесса было неотступное соблюдение закона, так как репутация компании могла рухнуть гораздо раньше, чем можно было бы воспользоваться экономическими преимуществами промышленного шпионажа. Вскоре эти методы работы начали применяться и другими американскими компаниями. Затем бизнес-разведка стала применяться в Европе, а в дальнейшем и во всем мире.

Игнорирование возможностей бизнес-разведки на начальном этапе дорого обходилось даже для крупнейших компаний [Джилад, 2010]. Так после создания фотоаппарата, который выдавал готовый снимок, компания Polaroid стала почивать на лаврах. Когда аналитический отдел компании представил отчет, в котором указал на перспективы развития фотоиндустрии и зарождении цифровой эры, руководство компании назвали эту информацию «футуристической чепу-

хой». Прошло некоторое время и в октябре 2001 года компания Polaroid начала первую процедуру банкротства.

Аналогично в 70-х годах XX века «Большая тройка» американских производителей автомобилей не прореагировала на появление на рынке японских производителей автомобилей. Однако, сами американцы выбрали небольшие, экономичные и надежные японские автомобили, и американские корпорации понесли значительные убытки.

Бизнес-разведчики из корпорации Samsung узнали из открытой прессы, что последний американский завод по производству гитар может закрыться из-за более дешевых корейских инструментов, и американское правительство готовится защитить своих производителей с помощью таможенных пошлин. Вовремя узнав это, представители Samsung успел ввезти в США большое количество гитар, а в результате введения ввозных пошлин, еще и поднять цены на этот музыкальный инструмент.

Сегодняшнее развитие информационных технологий сделало компьютерную конкурентную разведку доступной даже для относительно небольших компаний, сегодня она распространена на всех уровнях экономики.

На практике понятийная база конкурентной разведки еще окончательно не сформировалась, пока не делается разницы между терминами «деловая» или «экономическая» разведка, и под конкурентной разведкой ошибочно понимают весь комплекс мероприятий, связанный с информационно-аналитическим обеспечением управления предпринимательскими рисками, выявления угроз, возможностей и других факторов, влияющих на получение конкурентных преимуществ в бизнесе.

В арсенале тех, кто сегодня полноценно занимается конкурентной разведкой, нет специальной аппаратуры, шпионской техники. Их основной инструмент – компьютер, подключенный к сети Интернет. Деятельность подразделений (служб) конкурентной разведки компаний все больше основывается на последних достижениях в области искусственного интеллекта в сочетании с наработками в областях психологии, социологии, экономики.

Ощутимые преимущества, получаемые за счет использования конкурентной разведки, подтверждают результаты

опроса, проведенного еще в 1999 г. среди 500 крупнейших компаний США. Почти 90 % компаний подтвердили, что создали у себя подразделения конкурентной разведки. При этом затраты корпораций на разведку составляют в среднем 1–1,5% от оборота и вполне рентабельны [Ландэ, Прищепа, 2007].

В настоящее время создаются многочисленные профессиональные объединения (сообщества) специалистов в области конкурентной разведки. Наиболее известные из таких сообществ, занимающихся организацией конференций, тренингов, – это Strategic and Competitive Intelligence Professionals, SCIP в США (www.scip.org) (Рис. 1) и Competia в Канаде (www.competia.com).

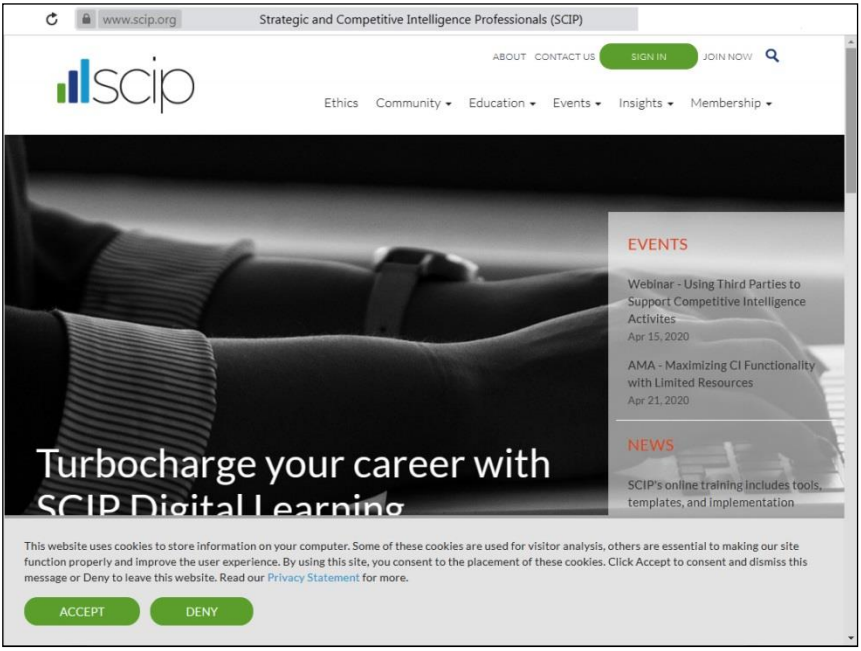


Рис. 1. Фрагмент веб-сайта организации SCIP (www.scip.org)

В Украине известна общественная организация «Общество аналитиков и профессионалов конкурентной разведки» (https://www.scip.org.ua/). В Украине ведется подготовка

специалистов в области конкурентной разведки в Харьковском национальном университете радиоэлектроники, где готовят магистров по специальности «Консолидированная информация».

С началом российской агрессии в отношении Украины вопросы получения информации из открытых источников и развенчание фейков приобрели новое значение и актуальность. Это дало старт новым проектам и учебным платформам, целью которых стало формирование и распространение навыков и умений работать с информацией онлайн, проверять информацию и т.д. Одним из таких проектов стала OSINT Academy и бесплатный онлайн курс по OSINT (Open Source Intelligence) Института постиндустриального общества [OSINT, 2018].

В настоящее время конкурентная разведка не ограничивается изучением конкурентов, а проводит анализ всей среды, окружающей организацию или предприятие. Изучается политическая обстановка, особенности законодательства, кадровые перемещения, новые технологии, собственные клиенты и поставщики компании и т.п., подбираются эксперты по специальным вопросам.

1. Конкурентная разведка и OSINT

1.1. Задачи конкурентной разведки

Основными задачами интернет-разведки, как сегмента конкурентной разведки [Кочергов, 2009], являются:

1. Информационное обеспечение процесса выработки управленческих решений на стратегическом и тактическом уровнях;

2. Раннее предупреждение, т.е. привлечение внимание лиц, принимающих решения, к угрозам, которые потенциально могут причинить ущерб бизнесу;

3. Прогноз и предотвращение возможных угроз бизнесу;

4. Выявление (совместно со службой безопасности) попыток конкурентов получить доступ к закрытой информации компании.

5. Определение благоприятных возможностей для бизнеса;

6. Управление рисками, обеспечение эффективного реагирования компании на быстрые изменения окружающей среды, интернет-пространства;

7. Промышленная контрразведка, упреждение разведывательной деятельности конкурентов в сетевой среде, аналитическая поддержка службы безопасности компании.

Приведенные выше задачи конкурентной разведки являются ключевыми, они служат достижению фундаментальной цели конкурентной разведки — обеспечить защищенность компании, осознанию того факта, что ее судьба находится в руках лиц принимающих решения, что компания не станет внезапно жертвой чьих-то враждебных действий.

Кроме того, в рамках компьютерной конкурентной разведки должны быть решены следующие задачи:

- сбор и своевременное обеспечение руководства и бизнес-подразделений компании надежной и всесторонней информацией из сетевых источников о «внешней» и «внутренней» среде предприятия;
- выявление факторов риска, угроз, которые могут затронуть экономические интересы бизнеса или помешать его нормальному функционированию;

- выявление новых возможностей и других факторов, влияющих на получение конкурентных преимуществ;
- усиление благоприятных и локализация неблагоприятных факторов конкурентной среды на деятельность бизнес-структуры;
- выработка прогнозов и рекомендаций по влиянию конкурентной среды на деятельность бизнес-структуры.

Конкурентная разведка становится современным направлением исследования поведения конкурентов на рынке, позволяющим создавать модели рынка, его участников, определения характеристик и оптимизации тактики и стратегии развития субъектов хозяйствования на определенных рынках. Для решения ее задач требуется использование эффективных приемов работы с информацией и ее элементами. Информация при этом становится как объектом исследования рынка, так и основой для создания его модели.

Выше сформулированы задачи компьютерной конкурентной разведки, рассчитанные на легитимную деятельность соответствующих структур. Вся система конкурентной разведки должна позволять руководству компании не только оперативно реагировать на изменения ситуации на рынках, но и оценивать дальнейшие возможности своего развития. Конкурентная разведка обеспечивает переход от традиционного принятия решений на основе недостаточной информации к управлению, основанному на знаниях. При этом она также обеспечивает снижение рисков, безопасность бизнеса, а также приобретение конкурентных преимуществ. Современная система конкурентной разведки позволяет не только осуществлять мониторинг информации, но и моделировать стратегию конкурентов, выявлять их партнеров, поставщиков, понимать условия сотрудничества.

Основные задачи систем конкурентной разведки заключаются в нахождении и обобщении информации о конкурентах, рынках, товарах, бизнес-тенденциях и операциях по таким основным объектам:

- партнеры, акционеры, смежники, союзники, контрагенты, клиенты, конкуренты;

- объединения компаний, слияния, поглощения, кризисные ситуации и т.п.;
- кадровый состав компании, партнеров, конкурентов и т.д., а также кадровые изменения, их динамика;
- торговый оборот, бюджет и его распределения по пунктам;
- заключенные договора, соглашения или договоренности.

Интерес при проведении конкурентной разведки вызывает сфера деятельности компаний, сферы их влияния и интересов. Эти знания могут применяться, например, для оказания влияния на позиции партнеров и оппонентов. Большое значение имеет информация, относящаяся к политике конкурентов, их намерениям, сильным и слабым сторонам, продукции и услугам, ценам, рекламным кампаниям, другим параметрам рынка.

1.2. Особенности компьютерной конкурентной разведки

Современные открытые сетевые ресурсы, веб-сайты, социальные сети, видеосервисы, мессенджеры превращается в настоящее время в основной источник информаци и эффективный инструмент для конкурентной разведки. Они позволяют не только в режиме реального времени отслеживать действия компаний-конкурентов, но и выявлять последние тенденции по интересующей тематике. Назовем лишь некоторые способы использования интернет-ресурсов для решения задач конкурентной разведки [Lande, 2019]:

1. Получение новостей по целевой тематике.

Современные сетевые новостные сервисы, такие как Google News, Yahoo News, UAPort.net, социальные сети типа Twitter, FaceBook, Reddit позволяют получать новости, отобранные в соответствии с информационными потребностями пользователей. Например, при использовании социальной сети Twitter, можно воспользоваться поисковым режимом, и ввести запрос, например «банкротство». После этого пользователь получит список сообщений, в некоторых случаях снабженных аккаунтами пользователей, чьи сообщения релевант-

ны введенному запросу. Таким образом, можно определить экспертов, которых можно сгруппировать в соответствии со своими информационными потребностями. Затем, следуя за мнением группы экспертов, можно получить достаточно широкий охват проблемы, несколько точек зрения, новые информационные ресурсы.

2. Выявление тенденций.

По выбранным с помощью поисковых возможностей информационным ресурсам (веб-сайтам, социальным сетям и т.д.) можно вручную или с использованием специальных аналитических инструментов выявлять тенденции в выбранной сфере.

3. Получение рассылки целевых документов по подписке (сообщения мессенджеров, электронная почта, СМС).

Многие из новостных агрегаторов и социальных сетей (в частности, Twitter) предоставляют возможность качественных персонифицированных периодических рассылок, охватывающих сообщения, комментарии, экспертные каналы.

4. Построение сетей информационных связей, когнитивных карт

Для задач конкурентной разведки важно не только получение целевой информации (сообщений), но и понимание связей, которые обнаруживаются при анализе информации. Важен не только объект анализа, но и связанные с ним информационные ресурсы, профили в социальных сетях, «друзья», группы обсуждений и т.п. В некоторых случаях можно посмотреть, кто является подписчиком данных профилей, кто интересуется той же тематикой и, следовательно, может стать новым источником для получения целевой информации.

5. Получение ответов на вопросы.

Социальные сети, форумы, блоги можно использовать как способ получения ответов на конкретные вопросы, в том числе и по вопросам методологии конкурентной разведки. Если вопрос поставлен корректно, то с большой вероятностью можно получить ответ на него от других пользователей.

6. Фильтрация «мусора».

Для конкурентной разведки не всегда интересны общеизвестные, зачастую ложные данные и информация (Fake News), интересные большинству, а ведь именно на такие данные ориентированы социальные сети. При использовании

сетевых ресурсов в качестве мощнейшей базы для конкурентной информации особое внимание следует уделять отработке запросов, выбору источников, экспертов, установлению связей.

1.3. Проблемы компьютерной конкурентной разведки

Отметим некоторые проблемы, связанные с компьютерной конкурентной разведкой.

Первой и наиболее существенной проблемой является то, что колоссальные объемы информации в Интернет, в частности, в социальных сетях, затрудняют поиск и выбор действительно необходимых сведений. Сами по себе необработанные, необобщенные и непроверенные данные не могут обеспечить качественную поддержку при принятии решений в области конкурентной разведки.

В настоящее время поисковыми системами, в частности, системой Google индексируется свыше триллиона документов, объемы постоянно растут. Наряду с этим, по словам Эрика Шмидта (Eric Emerson Schmidt) – председателя совета директоров Google с 2001 по 2011 годы, даже такая мощная поисковая система как Google сможет проиндексировать всю имеющуюся сегодня информацию лишь примерно через 300 лет.

Традиционные поисковые системы в Интернет отлично справляются с простыми однократными запросами, однако, как правило, слабо применимы для нужд конкурентной разведки. По некоторым оценкам [Додонов, 2014], более 97 % критичной для конкурентной разведки открытой информации невозможно найти с помощью традиционных информационно-посковых систем.

Второй проблемой компьютерной конкурентной разведки является то, что информация в Интернет имеет динамичный характер: она размещается, модифицируется и удаляется. Частичное решение этих проблем возможно при применении систем контент-мониторинга информационных потоков в Интернет.

Третья проблема, которую необходимо решить в целях конкурентной разведки, – автоматическое извлечение понятий из формализованных массивов информации (таблиц, баз

данных), а также неструктурированных текстов. Перспективным направлением решения этой проблемы в системах конкурентной разведки является использование технологий Knowledge Discovery, Data Mining и Text Mining [Ландэ, 2005, Надо новое], [Печенкин, 2004].

Четвертой проблемой является возможность автоматического выявления неочевидных закономерностей и связей, зафиксированных в документах. В настоящее время известно несколько путей решения проблем извлечения понятий из текстов и выявления их взаимосвязей, как практических, так и теоретических. Одним из этих путей является построение матриц и графов связей понятий, моделей предметных областей, когнитивных карт, к которым можно применять соответствующие математические методы. Как правило, узлы этих графов – коэффициенты, которые пропорциональны количеству документов, соответствующим исследуемым понятиям.

Пятой проблемой является поиск информации в «скрытом» Интернете, где содержится несравнимо большее количество данных, потенциально интересных для конкурентной разведки, чем в открытой части сети. Не вся потенциально открытая «несекретная» информация является хорошо доступной, скорее – наоборот. Добыча необходимой в каждом конкретном случае информации является сложной задачей. По мнению экспертов, только порядка 10-15% необходимой информации имеется в Интернете в готовом виде, остальные 85-90% можно получить в результате сравнения, агрегации и анализа многочисленных разрозненных данных.

Итак, в Интернет содержится большая часть информации, необходимой для проведения конкурентной разведки, однако остается открытым вопрос ее нахождения и эффективного использования. Причина – присущие сети Интернет недостатки [Ландэ, 2005]:

- непропорциональный рост уровня информационного шума;
- засилье паразитной информации;
- слабая структурированность и связность информации;
- динамичность информации;
- отсутствие целостности информации;

- многократное дублирование информации;
- отсутствие возможности смыслового поиска;
- ограниченность доступа к «скрытому» веб.

Несмотря на это возможности Интернета оцениваются экспертами в области конкурентной разведки достаточно высоко.

1.4. OSINT – разведка по открытым источникам

1.4.1. OSINT как дисциплина разведки

В качестве одного из синонимов понятия конкурентной разведки, часто используемого в силовых ведомствах различных государств, используется понятие «разведки по открытым источникам» OSINT (Open source intelligence). Это одно из направлений разведки, которое включает в себя поиск, выбор и добывание разведывательной информации, полученной из общедоступных источников (не обязательно компьютерных или сетевых), а также анализ этой информации.

OSINT базируется на двух основных понятиях:

- открытый источник – это источник информации, который предоставляет ее без требования сохранения ее конфиденциальности, т.е. предоставляет информацию, не защищенную от публичного раскрытия. Открытые источники относятся к среде общедоступной информации, и не имеют ограничения в доступе для физических лиц;
- общедоступная информация – это информация, опубликованная или размещенная для широкого использования; доступная для общественности.

По утверждениям аналитика ЦРУ Шермана Кента на 1947 год, политики получают из открытых источников до 80 процентов информации, необходимой им для принятия решений в мирное время. Позднее генерал-лейтенант Самуэль Уилсон, руководителя Разведывательного управления Министерства обороны США в 1976—1977 годах, отмечал, что «90

процентов разведданных приходит из открытых источников и только 10 — за счёт работы агентуры».

Американский исследователь по вопросам безопасности Марк М. Ловенталь определяет открытую информацию как «любую информацию, которая может быть получена из открытых коллекций: все типы СМИ, правительственные отчеты и другие документы, научные исследования и отчеты, коммерческие поставщики информации, Интернет и т. д. Основная характеристика открытой информации – это то, что для ее получения не требуются нелегальных методов сбора и что она может быть получена с помощью средств, полностью соответствующих авторским правам и коммерческим условиям поставщиков.

Мировое сообщество все больше использует информацию из открытых источников в целях решения широкого спектра задач. Материалы OSINT служат базой для всех методов ведения разведки как накопитель разведывательных данных, их анализатор и распространитель.

В соответствии с [ATP, 2012] разведка в открытых источниках OSINT является одним из способов ведения разведки, который вносит значительный вклад при планировании боевых действий, а также предоставляет всю необходимую информацию при их проведении. Также определяется:

1) Разведка в открытых источниках (OSINT) является одним из методов ведения разведки путем сбора информации из открытых источников, ее анализа, подготовки и своевременного предоставления конечного продукта вышестоящему руководству в целях решения определенных разведывательных задач.

2) OSINT является методом ведения разведки, разработанным на основе сбора и анализа общедоступной информации, и не находящимся под непосредственным контролем правительства США. OSINT является результатом систематизированного сбора, обработки и анализа необходимой общедоступной информации.

В частности, роль OSINT при проведении разведки определяется рядом аспектов, среди которых оперативность поступления, объем, качество, ясность, легкость дальнейшего использования, стоимость получения и т.д. Следующие фак-

торы влияют на процесс планирования и подготовки ведения OSINT:

- Эффективное информационное обеспечение. Большая часть необходимых справочных материалов об объектах информационных операций добывается из открытых источников. Это в основном достигается путем сбора информации из СМИ. Накопление данных из открытых источников является основной функцией OSINT.
- Релевантность. Доступность, глубина и масштабы публично доступной информации позволяют находить необходимую информацию без привлечения специализированных человеческих и технических средств разведки.
- Упрощение процессов добывания данных. OSINT предоставляет необходимую информацию, исключая потребность в привлечении излишних технических и человеческих методов ведения разведки.
- Глубина анализа данных. Являясь официальной частью разведывательного процесса, OSINT позволяет руководству осуществлять глубокий анализ общедоступной информации в целях принятия соответствующих решений.
- Оперативность. Резкое сокращение времени доступа к информации в сети Интернет. Сокращение человеко-часов, связанных с поиском информации, людей и их взаимоотношений на основе открытых источников. Быстрое получение ценной оперативной информации. Стремительно меняющаяся обстановка во время кризисов полнее всего отражается в текущих репортажах CNN с места событий.
- Объем. Возможность массового мониторинга определенных источников информации, с целью поиска интересующего контента, людей и событий. Как показывает опыт, грамотно собранные фрагменты информации из открытых источников в совокупности могут быть эквивалентны или даже более значимы, чем профессиональные разведывательные отчеты.

- Качество. По сравнению с отчетами специальных агентов информация из открытых источников оказывается предпочтительнее уже потому, что лишена субъективизма, не разбавлена ложью.
- Ясность. Так что если в случае использования OSINT надежность открытых источников бывает как ясной, так и неясной, то в случае с тайно добытыми данными степень их надежности всегда вызывает сомнения.
- Легкость использования. Любые тайны принято окружать барьерами из грифов секретности, особых режимов доступа. Что же касается данных OSINT, то их можно легко передавать в любые заинтересованные инстанции. Возможно проведение комплексного расследования на основании данных из Интернета
- Стоимость. Стоимость добычи данных в OSINT минимальна, определяется лишь стоимостью используемого сервиса.

В частности, сегодня, предлагаемые для OSINT программно-технологические решения обеспечивают:

- сбор данных из социальных сетей, таких как Facebook, Twitter или Youtube, анализ собранных данных;
- экстрагирование из собранного контента сути событий;
- агрегирование информации, полученной из сети Интернет;
- информационное влияние в сети Интернет;
- оценку достоверности информации;
- мониторинг и распознавание идентичности в сети Интернет, в том числе с помощью геолокации;
- работу с информацией, полученной из невидимых с помощью традиционных сетевых поисковых систем сегментов веб-пространства (dark web, hidden web, deep web).

1.4.2. Области применения OSINT

Существует множество применений OSINT, среди которых можно назвать:

Разведка

Открытые источники содержат огромное количество информации, необходимой и удовлетворяющей требованиям разведывательных органов, как государственных, так и частных, коммерческих, обеспечивающей понимание объективных и субъективных факторов, связанных, например, с деятельностью конкурентов. При этом, безусловно, для повышения эффективности разведывательной деятельности открытая информация используется в комплексе с другими, в частности, агентурными ресурсами.

Инициатива Американского сообщества разведки по открытым источникам (известная как National Open Source Enterprise) выражена Директивой сообщества разведки 301, которая обнародована директором Национальной разведки [DNI, 2006]. Директива устанавливает полномочия и обязанности помощника заместителя директора Национальной разведки по открытым источникам (ADDNI/OS), Центра открытых источников DNI и Национального комитета по работе с открытыми источниками.

OSINT в вооруженных силах

Ниже в качестве примера приведены подразделения вооруженных силы США, которые участвуют в деятельности OSINT:

- Unified Combatant Command;
- Defense Intelligence Agency;
- National Geospatial-Intelligence Agency;
- US Army Foreign Military Studies Office;
- EUCOM JAC Molesworth;
- Foreign Media Monitoring in Support of Information Operations, U.S. Strategic Command.

Национальная безопасность

Министерство внутренней безопасности (Department of Homeland Security) США включает активное разведыватель-

ное подразделение для работы с открытыми источниками. 14 февраля 2007 г. был учрежден «Domestic Open Source Enterprise» для поддержки департамента OSINT и работы с государственными, местными и племенными партнерами.

Юстиция

Сообщество правоохранительных органов OSINT применяет разведку по открытым источникам для прогнозирования, предотвращения, расследования преступлений и преследования преступников, включая террористов. Кроме того, центры центры обработки информации (Fusion Centers) в США все чаще используют OSINT для поддержки их разведки и расследований. Такие центры первоначально создавались под эгидой Министерства национальной безопасности (DHS) и Минюста и позволяли обмениваться стратегической информацией между ЦРУ, ФБР, Минобороны, МЧС, а также локальными администрациями и т.д.

Примерами успешных правоохранительных органов OSINT являются Scotland Yard OSINT; Королевская канадская конная полиция (RCMP) OSINT.

Полицейский отдел Нью-Йорка (NYPD) включает подразделение OSINT как отдел шерифа округа Лос-Анджелес, расположенный в Бюро по чрезвычайным операциям и связанный с Объединенным региональным разведывательным центром Лос-Анджелеса.

В плане правоохранительной деятельности OSINT может применяться при борьбе с такими явлениями, как:

- Организованная преступность и банды;
- Педофилия;
- Кража персональных данных и вымогательство;
- Отмывание денег;
- Преступность в сфере нарушения интеллектуальной собственности;
- Деятельность экстремистских организаций.

При этом с помощью OSINT обеспечивается выявление вовлеченности и усиление влияния в Интернете:

- Идентификация ключевых фигур и активистов;

- Мониторинг конкурента в режиме реального времени;
- Ограничение распространения информации;
- Формирование общественного мнения;
- Выявление экстремистских организаций;
- Риски для общественного транспорта;
- Санкции и правовые требования;
- Анализ баз данных противников (HME, IED, TTPs);
- Геолокация целей;
- Поддержка для военных операций.

Кибернетическая безопасность

В рамках OSINT обеспечивается поддержка процессов обеспечения кибербезопасности, в частности, могут быть даны ответы на такие вопросы из области защиты телекоммуникационных сетей посредством получения информации:

- Кто атакует вашу организацию?
- Каковы их мотивы?
- Как они организованы?
- Какие инструменты используют?

Бизнес

OSINT в бизнесе включает в себя коммерческую разведку, интеллектуальную разведку и бизнес-аналитику, и часто является основной областью практики частных разведывательных агентств.

Предприятия могут использовать информационных брокеров и частных следователей для сбора и анализа соответствующей информации для деловых целей, которые могут включать в себя средства массовой информации, глубокую сеть, веб нового поколения и коммерческий контент.

1.4.3. Технологии OSINT

OSINT – это очень разнообразная форма сбора и анализа информации. При работе OSINT часто следует принимать меры предосторожности при сборе информации из Интернета. Это может быть сделано в форме использования VPN для анонимности и незаметного сбора информации, прокси-серверов в распределенной сетевой среде. Оценка источни-

ков становится важной для общего процесса сбора и анализа OSINT. Аналитик OSINT нуждается в интеллектуальном анализе для выявления истинного или ложного процесса, который повлияет на прогнозирование будущего. Наконец, аналитики должны найти использование оценочного анализа для того, чтобы его результаты можно было включить в готовый классифицированный, неклассифицированный или запатентованный интеллектуальный продукт.

Сбор информации в OSINT, как правило, отличается от сбора данных в других разведывательных дисциплинах, где получение необработанной информации, подлежащей анализу, может быть основной трудностью. В OSINT основной трудностью является определение релевантных, надежных источников из огромного количества общедоступной информации.

Этапы OSINT

Процесс OSINT состоит из четырех этапов: планирования, подготовки, сбора и производства конечного материала – аналитики и четырех основных процессов: анализа, добычи и накопления разведданных, оценки и распределения по направлениям. Процесс ведения разведки, также как и процессы подготовки ответных информационных операций (планирование, подготовка, выполнение и подведение итогов), пересекаются и повторяются в соответствии с требованиями практики.

Как указано в «Инструкции по ведению разведки в полевых условиях», OSINT повышает эффективность и оказывает поддержку процессу ведения разведки и других операций.

На Рис. 2 представлена типовая схема процесса ведения OSINT.

Сбор разведданных синхронизирует и интегрирует процессы планирования, использования сил и средств, обработки и распределения элементов системы для поддержки боевых операций, что является объединенной разведывательной и оперативной функцией.

После анализа, информация, полученная из различных источников, становится разведданными, которые содержат не-

обходимую информацию о противнике, угрозах, климате, погодных условиях, рельефе местности и др.

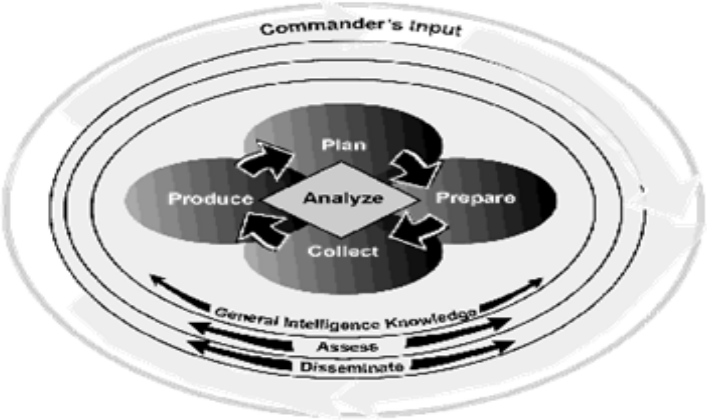


Рис. 2 – Типовая схема процесса ведения OSINT: План ⇒ Подготовка ⇒ Сбор ⇒ Производство. Общие знания разведки. Оценка. Распространение [APT, 2012]

Установлено, что такие элементы структуры OSINT, как постоянный поток информации, технические средства, программное обеспечение, безопасность средств коммуникации и базы данных охватывают средства:

- обеспечения доступности разведывательных данных. Обеспечение доступности разведданных является процессом, благодаря которому разведывательные организации активно и быстро получают доступ к разведданным;
- разработки и ведения автоматизированной сети разведки. Главной задачей является предоставление информационных систем, которые обеспечивают связь, совместный анализ и обработку, распространение материалов и создание условий доступности разведывательных данных;
- создания и поддержания доступа. Эта задача влечет за собой установление, обеспечение и поддержание доступа к секретным и несекретным программам, базам данных, сетям, системам и другим Интернет ресурсам для войск союзных государств, объеди-

ненных сил, национальных агентств и международных организаций;

- создания и ведения баз данных. Эта задача предполагает создание и поддержание несекретных и секретных баз данных. Создание и ведение базы данных способствует быстрому анализу, подготовке отчетов, обработке, распространению, ведению длительных боевых действий.

1.4.4. Международный опыт

Ведение разведки в открытых источниках повышает эффективность деятельности всего разведывательного сообщества, начиная национальным и заканчивая тактическим уровнями. Ниже приводится список некоторых организаций, которые занимаются в США добычей, накоплением, использованием, анализом, распространением информации из открытых источников.

- Совет по защите открытых источников (DOSC);
- Командование разведки и безопасности ВС США (INSCOM);
- Служба разведывательной информации Департамента сухопутных войск (DA IIS);
- Директор национальной разведки центра открытых источников (DNI OSC);
- Академия открытых источников;
- Департамент передовых систем (ASD);
- ФБР;
- Федеральный научно-исследовательский отдел (FRD), библиотека Конгресса.

Наряду с широким применением OSINT в США, приведем еще примеры применения этой технологии в других странах.

Служба внешней разведки Германии, Федеральная разведывательная служба, также использует преимущества Open Source Intelligence в подразделениях Abteilung Gesamtlage/FIZ и Unterstützende Fachdienste (GU).

В Австралии экспертом по открытым источникам является Управление национальных оценок (Office of National Assessments), которое является одной из разведывательных

госструктур. В Великобритании существует информационная служба BBC Monitoring, сосредоточенная на сборе открыто доступной информации силами журналистов. Анализом же собранных в BBC данных занимаются подписчики этого сервиса, в том числе и сотрудники секретных британских спецслужб.

2. Компьютерные технологии конкурентной разведки

Компьютерная конкурентная разведка использует в своем арсенале различные средства, наиболее развитыми из которых являются специализированные информационно-аналитические системы. Схема работы типовой информационно-аналитической системы компьютерной конкурентной разведки (ИАС ККР) приведена на Рис. 3.

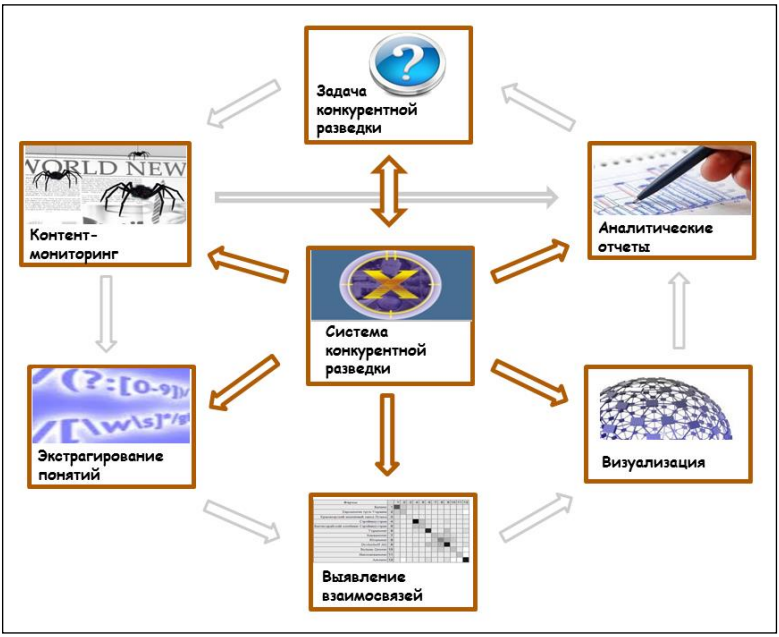


Рис. 3 – Схема работы типовой ИАС ККР

Информационно-аналитическая система компьютерной конкурентной разведки включает следующие компоненты:

- комплексы контент-мониторинга информации из открытых сетей (веб-пространства, социальных, пиринговых сетей и т.п.);

- средства экстрагирования понятий (компаний, персон, событий и т. п.) из полнотекстовых документов;
- средства выявления и визуализации информационных связей, выявления аномалий, неочевидных закономерностей;
- средства формирования аналитических документов, которые предоставляются лицам, принимающим решения (ЛПР).

Содержательная часть, информационная база информационно-аналитической системы конкурентной разведки формируется комплексом контент-мониторинга. Особенности современных комплексов контент-мониторинга заключаются в том, что они должны охватывать огромные объемы информации из динамически возрастающих информационных потоков в сетях при наличии шумовой информации, большой части слабодоступных ресурсов, так называемого «скрытого Интернета». В некоторых случаях реализация этого комплекса может быть передана так называемым «процессорам сбора данных», компаниям которые занимаются целенаправленным сбором больших объемов информации из социальных медиа по требованиям заказчиков.

Формирование базы данных (БД) ИАС ККР происходит путем подключения к сети Интернет и сбора (по определенным критериям и аккунтамы) информации из определенных информационных ресурсов (приведенный ниже список может расширяться):

Веб-сайты;
Блоги:
- Twitter;
- Livejournal.

Социальные сети:
- Facebook;
- Instagram;
- LinkedIN
- Reddit
- Medium.

Видеосервисы:
- YouTube;
- RuTube.

Мессенджеры:

Telegram;

Viber.

Кроме того, должна быть предусмотрена возможность настройки администратором комплекса контент-мониторинга ИАС ККР модулей автоматического сканирования и первичной обработки, а при необходимости создания служебных аккаунтов, через которые будет организован доступ к определенным информационным ресурсам.

С помощью комплексов контент-мониторинга в рамках конкурентной разведки, как правило, решаются следующие задачи:

- мониторинг деятельности партнеров, конкурентов, регулирующих органов;
- контроль медиаприсутствия и медиаактивности участников рынков;
- нахождение информации об участниках рынков;
- выявление новых продуктов на рынках;
- выявление новых игроков на рынках;
- организация ретроспективного информационного фонда документов для их последующего использования в аналитической деятельности.

Процесс превращения сырых данных в знания и доведение их до конечных потребителей принято называть разведывательным циклом. В своем классическом понимании разведывательный цикл (разведцикл) принято разделять на пять основных этапов:

- целеуказание, планирование, определение источников информации;
- сбор, добывание данных;
- обработка разведывательных данных (разведданных) – превращение их в разведывательную информацию;
- анализ и синтез разведывательной информации – превращение ее в знания – выводы, рекомендации, решения;
- доведение информации до конечных потребителей.

Следует также отметить некоторые ключевые особенности указанных выше этапов, а именно:

- целеуказание и планирование целесообразно делить на три уровня – стратегический, тактический и оперативный;
- на этапе сбора информации крайне важно задействовать как можно большее количество независимых и первичных источников;
- процесс обработки данных предполагает учет, классификацию, отбор, верификацию и оценку добытых сведений;
- разведовательный цикл, в некоторых случаях может не требовать глубокой проработки, например, в условиях ограниченного времени, может быть не полным и заканчиваться выдачей потребителям не знаний в виде окончательных выводов, рекомендаций или проектов решений, а просто обработанной информации в виде информационных справок;
- в разведывательном документе не должно быть ссылок на конфиденциальные источники информации, поскольку это может привести к их расшифровке;
- выводы и рекомендации должны быть четкими, краткими и однозначными, а прогнозы носить вероятностный характер;
- доведение информации до конечных потребителей должно быть в виде, адаптированном к восприятию заказчика и форме, легко доступной их пониманию (любопытно заметить, что ЦРУ, например, предоставляло президенту США Р. Рейгану ежедневную информацию в виде видеофильма, который снимали каждый день, поскольку бывший киноактер воспринимал такую подачу информации более адекватно).

Итак, открытые источники являются наиболее доступным каналом информации, при их использовании возрастает объективность добываемой информации, однако, резко возрастают трудозатраты на извлечение нужной информации. Следовательно, в компьютерной конкурентной разведке должны

применяться специализированные методики и системы. Такие специализированные методики и системы создавались учеными в интересах спецслужб на протяжении многих лет, как на Западе, так и в бывшем Советском Союзе. Перевод в последние 10–20 лет значительного объема мировой информации из бумажного вида в электронный, широкое использование и рост объемов сети Интернет, современные информационные технологии сделали конкурентную разведку в Интернете одним из самых перспективных направлений разведывательной деятельности. Тот факт, что так поступают практически все спецслужбы мира, лишь подтверждает перспективность этого направления.

Для поиска и сбора информации в компьютерных сетях в интересах разведки по всему миру используются специальные мониторинговые системы сбора данных, процессоры сбора данных, которые используют специальные программные комплексы (на компьютерном сленге их называют «роботами» или «пауками»). Программа-робот сама обходит по заданному графику указанные адреса (URL) в сети Интернет, скачивает с них данные, а затем извлекает из них нужную информацию, используя целый арсенал средств лингвистического, семантического и статистического анализа. Такие программные комплексы автоматически перехватывают любую поставленную на мониторинг информацию, как только она появится в доступном сегменте Сети.

При организации компьютерной конкурентной разведки широкое распространение получило использование направления науки, возникшего на стыке искусственного интеллекта, статистики и теории баз данных, как Knowledge Discovery (обнаружение знаний), использующего концепции Data Mining (глубинный анализ формализованных данных) и Text Mining/Information Extracting (глубинный анализ текстов/извлечение знаний из информации). Уникальными особенностями этих концепций и технологий является то, что с их помощью можно добывать из «сырых» данных ранее неизвестные, неочевидные, полезные на практике и доступные для интерпретации знания, необходимые для принятия решений в различных сферах деятельности. Такие технологии

применялись в основном в специальными службами. Одним из первых рассекреченных подобных комплексов стала французская система «TAIGA» (Traitement Automatique de l'Information Geopolitique d'Actualite – автоматическая система обработки актуальной геополитической информации) [Доронин, 2003]. Этот программный комплекс на протяжении 11 лет использовался в интересах французской разведки, после чего был заменен более современным, рассекречен и разрешен к коммерческому использованию. Новый более совершенный комплекс Noemic, поставленный на вооружение французской разведки, способен обрабатывать информацию со скоростью более 1 миллиарда знаков в секунду. Американский аналог этих программных комплексов Topic, что в переводе значит «Тема», также уже рассекречен и передан для коммерческого использования.

Аналогичные аналитические системы создавались в бывшем СССР, в частности в России. Достаточно вспомнить такие известные системы ФАПСИ, как «Барометр», «Эльбрус». Они занимались обработкой российской и зарубежной прессой, статистической и оперативной информацией.

Создание и использование таких систем продолжается. Так, например, система Radian 6 (www.radian6.com) предназначена для отслеживания в реальном времени упоминаний в социальных сетях брендов с учетом тональности и для участия в происходящих обсуждениях. Другая система – Alterian SM2, позволяет также отслеживать упоминания брендов, а также локализовать места обсуждений и определять демографические характеристики пользователей социальных сетей. По состоянию на 2021 год ведущими системами являются ActivTrak, ChartMogul, Cluvio, Databox, Matomo Analytics, Metabase, Tableau (https://blog.capterra.com/top-8-free-and-open-source-business-intelligence-software/). В Украине также созданы и развиваются десятки информационно-аналитических систем конкурентной разведки, речь о которых пойдет ниже.

На первый взгляд может показаться, что все перечисленные примеры – это системы, которые либо используются государственными структурами, либо слишком дороги, чтобы их

могли использовать «среднестатистические» компании. На самом деле все не совсем так. На современном рынке представлен целый ряд, как западных коммерческих продуктов, так и отечественных продуктов, способных в том или ином объеме выполнять подобные задачи в интересах конкурентной разведки коммерческих структур.

2.1. Поиск информации в Интернете

Для того, чтобы получить крупицы необходимой пользователю информации в Сети необходимо обработать огромные массивы сырых данных. Естественно, что для облегчения этой задачи используются специальные поисковые инструменты.

Поиск информации в сети Интернет только путем просмотра отдельных веб-сайтов, во-первых, носит выборочный и/или случайный характер (к тому же информация на отдельных сайтах может носить весьма субъективный или даже заказной характер), во-вторых, не продуктивен.

Все имеющиеся средства поиска информации в Интернете могут быть условно разделены на несколько подгрупп, а именно:

- средства поиска информации на отдельных сайтах;
- подборки ссылок, каталоги;
- поисковые системы;
- метапоисковые системы;
- системы мониторинга и контент-анализа;
- экстракторы объектов, событий и фактов;
- системы Knowledge Discovery, Data Mining, Text Mining;
- специализированные системы конкурентной разведки;
- интегрированные системы.

Все каталоги, поисковые системы и метапоисковые системы являются веб-сайтами со специализированными базами данных, в которых хранится информация о других веб-сайтах и документах, хранящихся на них. По запросу к таким системам выдается список гиперссылок, а иногда и краткое описание документов (сниппеты). Как правило, поиск может производиться по ключевым словам и фразам. Активизируя

на гиперссылку, найденную в результате запроса, пользователь попадает на оригинал документа. Естественно, что если документ со временем изменился или веб-сайт прекратил свое существование, то и первоначально заиндексированный поисковой системой документ через некоторое время может быть не найден.

Основное отличие поисковых систем от каталогов – наличие автоматического «робота», постоянно сканирующего вебпространство и накапливающего новую информацию в индексных файлах базы данных. В каталоги информация как правило заносится вручную – либо владельцами сайтов, либо обслуживающим персоналом самих каталогов. Пользование такими системами, как правило, бесплатное.

Метапоисковые системы являются системами, интегрирующими результаты поиска разными поисковым системам. Так как отдельные поисковые системы различным образом индексируют различные сегменты Сети, то, естественно, и результат поиска с помощью метапоисковой системы будет более полным, чем с помощью одной отдельно взятой поисковой системы. Вторым поисковым преимуществом таких систем является то, что одним запросом обеспечивается поиск во многих поисковых системах, не требуя многочисленных повторений одного и того же запроса.

Системы мониторинга обеспечивают регулярный поиск и «скачивание» информации по заданным темам и с заданных сайтов, а также анализ содержания «скачанных» документов. Такие системы, как правило, обладают развитым языком запросов, что позволяет существенно детализировать и конкретизировать запросы по сравнению с обычными поисковыми системами. Кроме того, такие системы хранят в своих базах данных полные тексты исходных документов, что обеспечивает сохранность этих документов во времени и возможность их обработки и контент-анализа, как в текущем времени, так и в ретроспективе. Существенным преимуществом таких систем является также то, что сложные запросы, состоящие из десятков или сотен поисковых слов и выражений, однажды составленные аналитиком-экспертом, могут быть сохранены в виде каталогизированного запроса или рубрики и в дальнейшем вызываться автоматически или вручную из сохраненного списка для проведения поиска или анализа.

С помощью контент-анализа такие системы позволяют устанавливать пересекающиеся связи между темами, понятиями и объектами, поставленными на мониторинг, выявлять эмоциональную окраску документов, проводить анализ динамики появления во времени тех или иных документов, проводить сравнительный анализ информационной активности по различным тематикам и многое другое.

Если мониторинговые системы как системы фильтрации могут выделять из информационного потока известные объекты, то экстракторы объектов, событий и фактов умеют выделять из потока информации объекты, неизвестные заранее, события или факты, которые лишь соответствуют определенному заранее типу, например, географические понятия, персоны, структуры и организации, события (дорожно-транспортные происшествия, катастрофы, международные встречи). При этом факты могут классифицироваться как обычные или необычные. Примером обычного факта в данном случае можно считать выезд автомобилей за черту города, а примером необычного факта – выезд за ту же городскую черту автомобиля без номерных знаков.

Системы типа Knowledge Discovery, технологии Data Mining и Text Mining, умеют выявлять новые знания и закономерности. Такая система, например, может самостоятельно, без участия человека, сделать вывод о факте знакомства между людьми, основываясь на имеющихся в системе данных об окончании ими одной и той же школы и одного итого же класса в одном и том же населенном пункте. Правда, сами правила, по которым такая система делает выводы, все-таки создаются и задаются пока что людьми.

Специализированные системы для конкурентной разведки могут включать в себя одно или несколько из перечисленных выше поисковых средств, приспособленных под эти специфические задачи. Кроме того, потребности конкурентной разведки предполагают использование в качестве источников информации, кроме полнотекстовых документов, еще и доступных в сети Интернет баз данных, собственных, принадлежащих компании, документов, таблиц и баз данных, а также формализованных и неформализованных документов и БД, добытых из других источников.

В странах Европейского союза обычный человек зарегистрирован в более 300 базах данных, таких как прописка (место жительства), страховка, водительские права, банки, кредитные бюро, информационные, рейтинговые, рекрутинговые агентства, бюро по трудоустройству, медицинские и полицейские учеты, супермаркеты, клубы, системы управления взаимоотношений с клиентами коммерческих фирм (так называемые CRM-системы) и т.п. В интересах конкурентной разведки и маркетинга анализируются не только рынки товаров и услуг, но и вкусы и предпочтения отдельных клиентов. Хранящаяся в различных базах данных информация о юридических лицах еще более обширна.

В целях бизнес-разведки необходимо анализировать данные из всех доступных источников информации, но в рамках данной работы не будут рассматриваться источники информации, не представленные в Интернете.

Интегрированные средства конкурентной разведки включают в себя не только все доступные поисковые средства, но и банк выявленных (добытых) и логично связанных между собой данных, информации и знаний.

С точки зрения создания информационно-аналитических систем такая система концептуально должна предполагать реализацию следующих трех принципов:

- единое информационное пространство взаимосвязанных концептов – объектов и фактов независимо от типа их источников или контента;
- сохранение связей концептов с релевантными данными и источниками информации;
- исторически-пространственная модель банка данных системы, которая предполагает наличие у всех объектов учета атрибутов времени и места.

Справедливости ради следует отметить, что, согласно отчетам Fuld's Intellgence Software Report, известных коммерческих версий полноценных интегрированных систем, позволяющих решать весь комплекс задач конкурентной разведки, пока не существует, по крайней мере, на Западе.

По мнению эксперта в области разведки А. Масаловича, из 23 видов поисковых задач, интересующих аналитика спецслужб традиционные поисковые системы удовлетворительно решают лишь одну. Поисковые системы отлично

справляются с простыми однократными запросами. Когда же предметная область сложна или слишком широка (например – «политика», «экономика»), или, наоборот, предельно узка и отдалена во времени (например, условия сделки некоторых компаний пятилетней давности), а требуется обобщить все информационные темы и поводы по данной тематике, оценить их во временной динамике, найти взаимосвязи с другими объектами, составить целостную картину об интересующем объекте, выделить нестандартное событие из общего массива, то можно убедиться, что:

- выдача поисковых систем либо перегружена тысячами бесполезных ссылок, либо наоборот недостаточна;
- информация в сети Интернет не хранится долго, необходимую информацию, присутствующую на целевом сайте месяц назад, сегодня можно там не обнаружить;
- поисковая система не сохраняет просмотренные аналитиком ссылки и ему каждый раз приходиться начинать рутинную работу с нуля после вынужденного перерыва;
- поисковая система не всегда отличает действительно важную информацию от информационного шума;
- поисковая система не всегда способна обобщать или сравнивать информацию по смыслу или другим содержательным критериям;
- поисковые системы не охватывают некоторые вебресурсы или отдельные виды информации (например, информацию из баз данных), а некоторые вебресурсы, наоборот, всегда показываются на первых страницах выдачи, хотя их содержание не интересно авторам запросов;
- поисковые системы могут выполнять поиск информации только по непосредственно введенному запросу и не всегда могут повторять их автоматически в заданное время без участия пользователя.

По оценкам экспертов [Кузнецов, 2006] значительную часть критичной для бизнеса информации из сети Интернет

невозможно найти с помощью традиционных информационно-поисковых систем. Точнее сетевые информационно-поисковые системы не в полной мере справляются с задачами конкурентной разведки. Поэтому разрабатываются специализированные системы, ориентированные на задачи сетевой аналитики, конкурентной разведки. Список таких общедоступных систем, например, приведен по адресу (http://hrazvedka.ru/category/poisk_soft). Приведем описание некоторых из них:

Website-Finder (www.softpedia.com) – программа, которая дает возможность поиска веб-сайтов, плохо индексируемых поисковой системой Google. Для каждого запроса выдается 30 результатов. Программа проста в использовании, есть бесплатна версия.

Global Supplier Directory by Solusource (www.worldindustrialreporter.com/solusource) – веб-интерфейс для конкурентной разведки от компании Thomas. Позволяет найти информацию, имеющуюся в ретроспективных базах данных Thomas (охват – более 100 лет) по компаниям, продуктам и отраслям.

dtSearch (www.dtsearch.com) – поисковая программа, позволяющая обрабатывать терабайты текста, как на локальном диске, так и в сетевом окружении. Поддерживает статические и динамические данные. Позволяет искать во всех форматах MS Office.

InfoNgen (www.infongen.com) – агрегатор, охватывающий в режиме просмотра свыше 35 тысяч онлайновых источников, легко настраиваемый на уникальные темы. Объединяет мониторинг, фильтрацию и агрегацию информации по запросам конкретного пользователя. Предоставляет информацию на восьми языках, обеспечивает перевод на английский язык.

Sentinel Vizualizer (www.fmsasg.com) – одна из лучших в мире программ по визуализации связей и отношений Sentinel Vizualizer.

Web Content Extractor (newprosoft.com) – "Web Content Extractor" является наиболее мощным, простым в использовании программное обеспечение извлечения данных из веб-сайтов.

Screen-Scraper (screen-scraper.com) – позволяет автоматически извлекать всю информацию с веб-страниц, скачи-

вать подавляющее большинство форматов файлов, автоматически вводить данные в различные формы. Работает под всеми основными платформами, имеет полнофункциональную бесплатную и очень мощные профессиональные версии.

Attackindex (attackindex.com) – система, позволяющая дать ответы на вопросы: ведется ли против пользователя информационная атака или произошел естественный всплеск интереса к событию; когда началась информ операция, насколько она интенсивна и масштабна; какие сайты и аккаунты в соцсетях используются для атаки; кто стал инициатором информационной операции и как связаны её участники (Рис. 4).

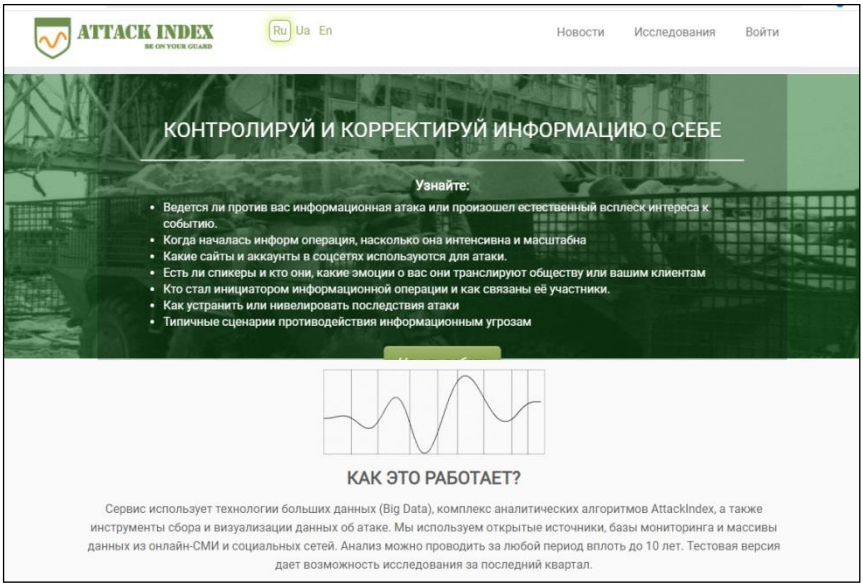


Рис. 4 – Фрагмент страницы сайта Attack Index (attackindex.com)

Photoinvestigator (photoinvestigator.co) – сервис для извлечения метаданных и другой информации из фотографий.

Visual.ly (visual.ly) – система поиска инфографики в вебпространстве.

CIRadar (www.ciradar.com/Competitive-Analysis.aspx) – коммерческая англоязычная система поиска информации для

конкурентной разведки в «глубинном» веб. Реализована как веб-сервис.

2.2. Мониторинг информационного пространства

Современные методы контент-мониторинга – это адаптация классических методов контент-анализа и глубинного анализа текстов (Text Mining) к условиям формирования и развития динамических информационных массивов, например, потоков информации из сети Интернет. Первая типовая задача контент-мониторинга – построение диаграмм динамики появления понятий (отражения событий) во времени.

На примере рынка нефтепродуктов рассмотрим, как из массивов текстовой информации из сети Интернет могут быть выявлены документы, содержащие максимальное количество ценовой информации по данному рынку.

Рассмотрим, как в системе контент-мониторинга InfoStream [Григорьев, 2007] отслеживаются публикации, относящиеся к российско-украинскому газовому кризису 2008–2009 годов. Для этого был составлен запрос ***«газов~криз & geo.UA»***, введенный через веб-интерфейс системы.

На появившейся после отработки запроса диаграмме видно, что пик кризиса пришелся на середину января 2009 года (Рис. 5) и был связан с подписанием соответствующего договора в Кремле и реакцией на это Секретариата Президента Украины (Рис. 6).

Кроме того, можно перейти в режим «Сюжеты», в котором предусмотрена кластеризация результатов поиска с учетом весовых критериев, что позволяет выдавать пользователю лишь наиболее весомые цепочки документов. Поэтому обеспечивается достаточно высокий уровень соответствия выдаваемых документов и потребности, выраженной запросом. Для получения списка основных сюжетов, относящихся к рынку нефтепродуктов, был введен запрос **«(нефтепродукт|бензин) & цены»**, который уточнялся специальными признаками **«numb.medium|numb.large»,** означающие в системе InfoStream средний или высокий уровень присутствия в документах цифровой информации (Рис. 7). После этого достаточно перейти в режим просмотра и проанализировать документы, ссылки на которые выданы системой (Рис. 8).

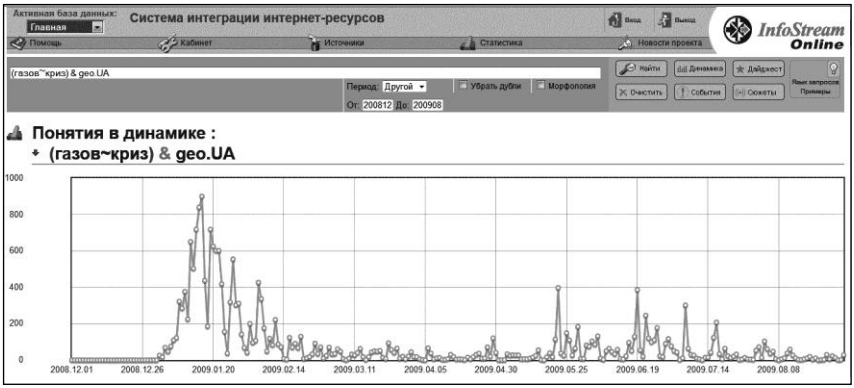


Рис. 5 – Диаграмма динамики понятия во времени

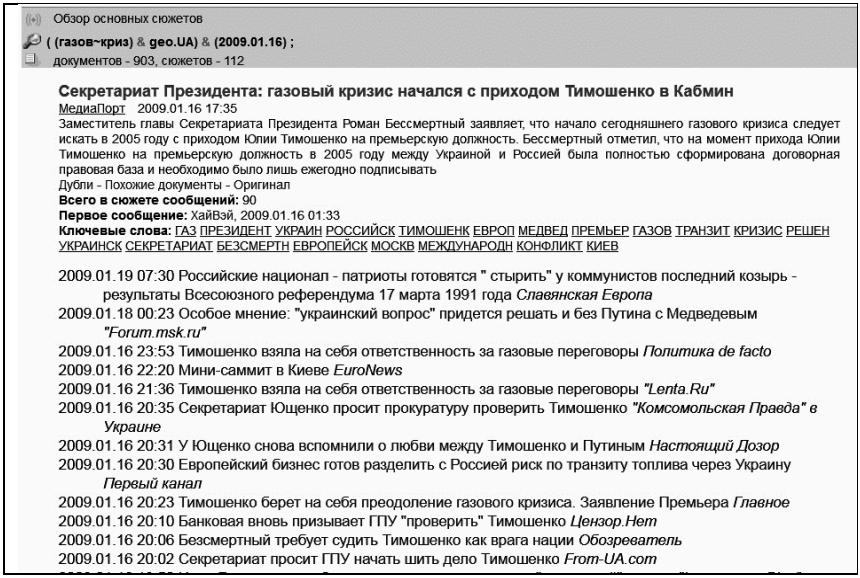


Рис. 6 – Основная сюжетная цепочка по запросу

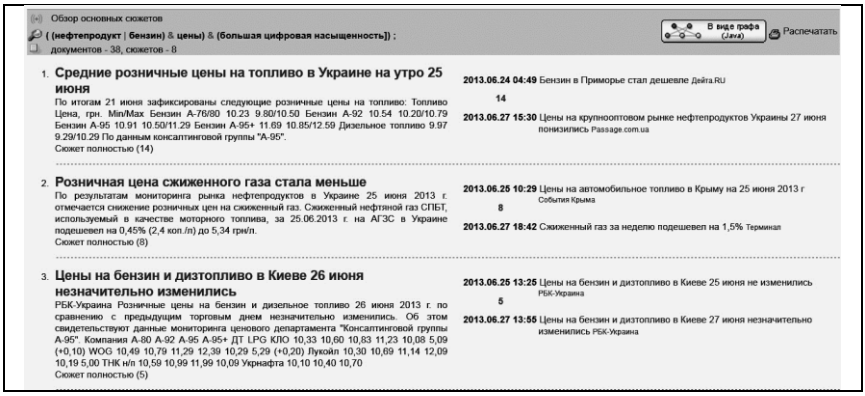


Рис. 7 – Фрагмент цепочки основных сюжетов

Цены на топливо на 27.06.2013

По данным консалтинговой группы "А-95", средние на бензин и дизельное топливо на АЗС в Днепропетровске, грн./л на 27 июня 2013 года:

Компания	А-80	А-92	А-95	ДТ (Л-02,62)
Укрнафта	10,10	10,40	10,70	9,80
Веста Сервис	10,10	10,34	10,59	9,59
Формула Ритейл		10,59	10,94	9,99
Лукойл		10,69	11,14	10,19
Нефтек	10,29	10,59	10,99	9,99
Альфа-Нафта		10,35	10,65	9,80
Средняя по области	**10,16**	**10,49**	**10,84**	**9,89**

Средние цены на топливо по областям Украины:

Область	А-80	А-92	А-95	ДТ (Л-02,62)
АР Крым	10,24	10,58	10,98	9,99
Винницкая	10,25	10,61	11,00	10,02
Волынская	10,23	10,54	10,90	9,95
Днепропетровская	10,24	10,45	10,77	9,85
Донецкая	10,25	10,70	11,16	10,14
Житомирская	10,29	10,49	10,85	9,95
Закарпатская	10,23	10,54	10,91	9,86
Запорожская	10,20	10,47	10,82	9,89
Ивано-Франковская	10,23	10,61	11,01	10,05
Киев	10,23	10,60	10,98	10,04
Кировоградская	10,24	10,53	10,90	9,96
Луганская	10,25	10,62	10,94	10,00
Львовская	10,23	10,60	11,00	10,05

Рис. 8 – Документ с ценовой информацией

2.3. Text Mining, Information Extraction

Задача, которую необходимо постоянно решать при проведении конкурентной разведки – автоматическое извлечение

понятий и фактов из формализованных массивов информации (таблиц, БД) и неструктурированных текстов, представленных в веб-пространстве, выявление глубинных связей между отдельными понятиями. Для этого предполагается использование в системах конкурентной разведки технологий Knowledge discovery, концепции глубинного анализа данных и текстов (Data Mining, Text Mining).

Важная задача технологии Text Mining связана с извлечением из текста его характерных элементов или свойств, которые могут использоваться в качестве метаданных документа, ключевых слов, аннотаций. Другая задача заключается в отнесении документа к некоторым категориям из заданной заранее схемы классификации. Text Mining также обеспечивает новый уровень семантического поиска документов.

Согласно сложившейся в настоящее время методологии, к основным элементам Text Mining относятся [Ландэ и др., 2009]: классификация (Classification), кластеризация (Clustering), построение семантических сетей, извлечение фактов, понятий (Feature Extraction), реферирование (Summarization), ответы на запросы (Question Answering), тематическое индексирование (Thematic Indexing) и поиск по ключевым словам (Keyword Searching). Также в некоторых случаях этот набор дополняется средствами поддержки и создания таксономии (Taxonomies), тезаурусов (Thesauri) и онтологий (Ontology).

При классификации текстов используются статистические корреляции для создания правил размещения документов в определенные категории. Задача классификации – это классическая задача распознавания, где по некоторой контрольной выборке система относит новый объект к той или иной категории. Особенность же концепции Text Mining заключается в том, что количество объектов и их атрибутов могут быть очень большими – предусматривается применение интеллектуальных механизмов оптимизации процесса классификации.

Кластеризация базируется на признаках документов, применении лингвистических и математических методов без использования заданных заранее категорий. Результатом кластеризации может быть таксономия или визуальная карта, которая обеспечивает эффективный охват больших объемов

данных. Кластеризация в Text Mining рассматривается как процесс выделения компактных подгрупп объектов с близкими свойствами. Средства кластеризации позволяют находить признаки и разделять объекты по подгруппам на базе этих признаков. Кластеризация, как правило, предшествует классификации, поскольку позволяет определить группы объектов.

При построении семантических сетей предполагается анализ связей между понятиями, экстрагируемыми из документов. Понятиям соответствует появление определенных дескрипторов (ключевых фраз) в документах. Связи между понятиями могут устанавливаться в простейшем случае путем учета статистики их совместного упоминания в различных документах.

Извлечение или экстрагирование фактов (понятий) предназначено для получения некоторых фактов из текста с целью улучшения классификации, поиска, кластеризации и построения семантических сетей.

Автоматическое реферирование (Automatic Text Summarization) [Хан, 2000] – это составление кратких изложений материалов, аннотаций или дайджестов, т.е. извлечение наиболее важных сведений из одного или нескольких документов и генерация на их основе лаконичных, понятных и информационно наполненных отчетов.

На основе методов автоматического реферирования возможно формирование поисковых образов документов. По автоматически построенным аннотациям больших текстов – поисковым образам документов – может проводиться поиск, характеризуемый высокой точностью (естественно, за счет полноты). В некоторых случаях вместо поиска в полных текстах массива больших по размеру документов оказывается целесообразным поиск в массиве специально созданных аннотаций. Хотя поисковые образы документов часто оказываются образованиями, лишь отдаленно напоминающими исходный текст, не всегда воспринимаемый человеком, но за счет вхождения наиболее весомых ключевых слов и фраз, они помогают приводить к вполне адекватным результатам при проведении полнотекстового поиска.

Уникальными особенностями концепции и технологий Text Mining, является то, что с их помощью можно извлекать

из «сырых» данных неочевидные, полезные на практике и доступные для интерпретации знания, необходимые для принятия решений в различных сферах деятельности, в том числе в области экономической конкуренции.

На современном рынке представлен целый ряд, как западных продуктов, так и систем производства постсоветских стран, способных в той или иной объеме осуществлять глубокий анализ текстов.

В последнее время все основные западные бренды, специализирующиеся на разработке информационных хранилищ и баз данных, корпоративных систем управления расширили свои линейки продуктов системами или модулями Text Mining. О наличии таких модулей заявляют SAP, Oracle, SAS, IBM и другие компании.

Процесс конкурентной разведки можно рассматривать как построение сети из исследуемых объектов и связей меду ними. Результаты должны представлять собой аналитическую информацию, которая может быть использована для принятия решений. Аналитическая информация может быть представлена в виде наглядных схем – семантических сетей, дайджестов, наборов сюжетных линий, взаимосвязей ключевых понятий, компаний, лиц, технологий и т.п.

Задачи конкурентной разведки породили спрос на специальные информационные технологии, обеспечивающие возможность извлечения и обработки необходимой информации, что в свою очередь вызвало поток предложений систем со стороны разработчиков программного обеспечения.

Сегодня решать задачи конкурентной разведки на основе информации из сети Интернет помогают общедоступные и специальные программы и сервисы, например, в последнее время приобрели популярность так называемые «персонализированные разведпорталы», способные отбирать информацию по самым узким, специфическим вопросам и темам и предоставлять ее заказчикам.

В настоящее время декларированы технологии и системы «компьютерной конкурентной разведки», идея которых заключается в автоматизации и ускорении процессов извлечения необходимой для конкурентной борьбы информации из открытых источников и ее аналитической обработки.

При ведении конкурентной разведки находят все более широкое применение новые направления науки и технологий, получившие названия: «управления знаниями» (Knowledge Management) и «обнаружение знаний в базах данных» (Knowledge Discovery in Databases) или иначе, Data и Text Mining – «глубинный анализ данных или текстов».

Если системы управления знаниями реализуют идею сбора и накопления всей доступной информации, как из внутренних, так и из внешних источников, то Data и Text Mining, как уже было показано, позволяют выявлять неочевидные закономерности в данных или текстах – так называемые латентные (скрытые) знания. В целом эти технологии еще определяют как процесс обнаружения в «сырых» данных ранее неизвестных, но полезных знаний, необходимых для принятия решений. Системы этого класса позволяют осуществлять анализ больших массивов документов и формировать предметные указатели понятий и тем, освещенных в этих документах.

Характерная задача конкурентной разведки, обычно включаемая в системы Text Mining – это нахождение исключений, то есть поиск объектов, которые своими характеристиками сильно выделяются из общей массы.

Еще один класс важных задач, решаемых в рамках технологии Text Mining – это моделирование данных, ситуационный и сценарный анализ, а также прогноз [Ланде, Фурашев, 2012].

Для обработки и интерпретации результатов Text Mining большое значение имеет визуализация. Часто руководитель компании не всегда адекватно воспринимает предлагаемую ему аналитическую информацию, особенно если она не вполне совпадает с его пониманием ситуации. В связи с этим служба конкурентной разведки должна стремиться представлять информацию в виде, адаптированном к индивидуальному восприятию заказчика.

Визуализация обычно используется как средство представления контента всего массива документов, а также для реализации навигации по семантическим сетям при исследовании, как отдельных документов, так и их классов.

2.4. Модели предметных областей

Важной задачей конкурентной разведки является выявление неочевидных закономерностей и связей из текстов вебстраниц и выявление их взаимосвязей, построение матриц и графов взаимосвязей.

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

Сегодня, когда практически у всех заинтересованных пользователей уже накоплен большой опыт работы с традиционными информационно-поисковыми системами, оказалось очевидным, что факты или понятия, которые ищутся с помощью таких систем, сами по себе зачастую бессмысленны. Например, если пользователя интересуют информационные связи Ощадбанка с другими банками или частными лицами, то он не знает, какие банки или фамилии ему указать в запросе, а все документы, содержащие слово «Ощадбанк», указать физически невозможно. В таких случаях информационные связи, количество которых выходит за рамки статистического фона, как правило, отражают реальность.

Интерпретируют обычно не сами понятия или факты, а взаимосвязи между ними. Важным оказывается не столько исследование самих понятий, сколько исследование их взаимосвязи. Известно, что именно взаимосвязь способствует пониманию мотивационно-целевых особенностей, то есть пользователя интересует не понятие само по себе, а понятие в окружении, чтобы сразу иметь представление о предметной области, при необходимости направить уточняющий поиск в нужном направлении. Подобные решения, реализованные в виде «информационных портретов», содержащих опорные слова, используются в таких системах, как InfoStream (infostream.ua), CyberAggregator.

База данных практически любой традиционной информационно-поисковой системы может рассматриваться в виде графа, вершинами которого выступают объекты – термы,

понятия, дескрипторы и др., а ребрами – их связи. Вместе с тем, основа поиска в этих случаях – поиск вершин, то есть поиск объектов. Поиск по взаимосвязям, ребрам, кажется на первый взгляд менее эффективным. Действительно, если предположить, что в графе N вершин, то число ребер теоретически может составлять $N(N-1)/2$, то есть, если предположить, что вершин всего 100 тыс., то ребер может оказаться около 5 млрд., что соответствует достаточно большой базе данных даже по современным понятиям. Вместе с тем, если в качестве вершин графа использовать такие понятия, как имена людей и названия компаний из новостных документов, то оказывается, что соответствующая матрица инцидентности оказывается очень разреженной. Измерения показали, что при количестве отдельных понятий, извлеченных из 5 млн. новостных документов, равном примерно N = 1,5 млн., количество связей составило всего лишь υ = 4 млн.

Кроме того, как показали эксперименты, распределение степеней вершин (степень вершины – количество исходящих из нее ребер) в подобных графах – степенное, что свидетельствует о, так называемой, безмасштабности, то есть о том, что многие характеристики (в частности, соотношение количества вершин и ребер), должно оставаться на одном уровне. Поэтому в качестве основы построения базы данных связей оказывается технически возможным использование ребер рассматриваемого графа – связей между отдельными понятиями.

В качестве массивов документальной информации для такой системы могут использоваться данные, поступающие от систем контент-мониторинга, таких как InfoStream, Webscan или «Яндекс.Новости» а также результаты мониторинга специализированных веб-служб, таких как базы данных биографий людей, организаций, служб трудоустройства и т.п.

Информационные взаимосвязи между понятиями выявляются путем обработки документальных массивов и, могут храниться в специальной базе данных. Набор понятий, используемый при построении базы данных связей, формируется путем экстрагирования данных из доступного пользователю текстового массива, что придает системе целостность.

В корпоративной информационной инфраструктуре база данных связей может использоваться различным образом,

например, отдельно, либо ее возможности могут быть дополнены возможностями существующих полнотекстовых и/или фактографических баз данных (Рис. 9). При этом основным результатом работы является построение так называемых «карт связей», а в качестве побочного эффекта, реализующего «режим доказательства», может рассматриваться извлечение самих документов как источников связей.

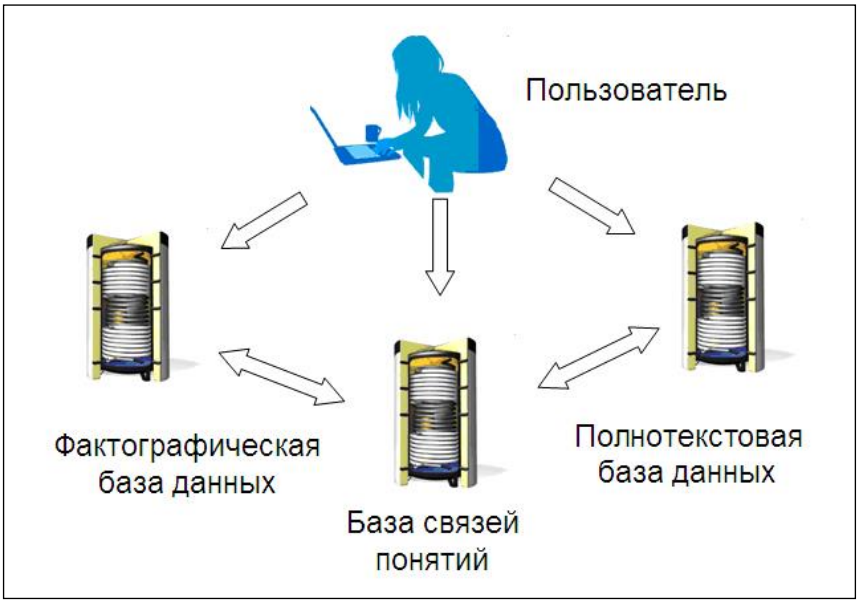


Рис. 9 – Место базы данных связей понятий
в информационной инфраструктуре

При проектировании баз данных связей используются перспективные решения в области создания информационно-аналитических систем, в частности, теория и технологии глубинного анализа тестов – Text Mining, в том числе методы экстрагирования информации (Information Extraction), технологии баз данных сверхбольших объемов (Big Data), концепция «сложных сетей» (Complex Networks).

В рамках теории сложных сетей изучаются характеристики, связанные с топологией сетей, но и статистические феномены, распределение весов отдельных вершин (в каче-

стве которых можно рассматривать сущности, понятия, факты) и ребер, эффекты протекания и проводимости в сетях и т.п.

На Рис. 10 схематически представлены возможные технологические этапы формирования базы данных связей [Ландэ, Брайчевский, 2010].

С помощью программы-робота осуществляется сканирование выбранных веб-ресурсов, содержащих информацию, относящуюся к объектам исследований.

После этого осуществляется экстрагирование необходимых пользователям понятий, например, наименований брендов, компаний, электронных адресов и т.п.

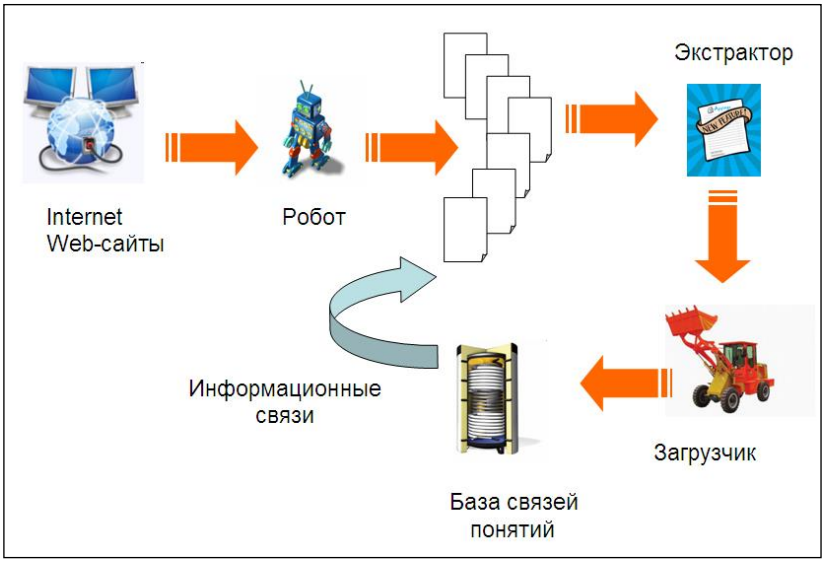


Рис. 10 – Схема формирования базы данных связей

Отобранные понятия и соответствующие отношения между ними загружаются в базу данных связей, которая также содержит ссылки на документы-первоисточники. Средства экстрагирования понятий, как правило, ориентированы на обработку документов, сканируемых из сети Интернет, представленных на различных языках.

Предложенный подход к поиску, естественно, влечет за собой некоторые особенности в реализации архитектуры базы данных связей понятий. В настоящее время наиболее популярной платформой для такой базы данных является графовая СУБД Neo4j. Кроме того, архитектура базы данных связей должна быть ориентирована на такие возможные применения, как выявление неявных связей (не выявленных явно комплексом экстрагирования понятий), поиск отдельных объектов, а также взаимосвязь с существующими фактографическими базами данных.

Можно назвать несколько систем, в которых частично реализован данный подход:

PolyAnalyst (www.megaputer.ru) – позволяет решать проблемы прогнозирования, классификации, группирования объектов, проводить анализ связей, многомерный анализ и интерактивное создание отчетов. Система PolyAnalyst (и ее компонента – система TextAnalyst) обеспечивает лингвистический и семантический анализ текста, выявление сущности, визуализацию связей, систематизацию документов, резюмирование и обработку запросов на естественном языке;

Sap Businessobjects Text Analysis (https://www.sap.com/sapbusinessobjects) – программа, позволяющая извлекать информацию о десятках типах объектов и событий, включая людей, географические названия (топонимы), компании, даты, денежные суммы, email-адреса и выявлять связи между ними;

Neticle Text Analysis (https://neticle.com/textanalysisapi/) – технология извлечения информации из неструктурированных текстов. Она позволяет выявлять информацию, содержащуюся в неструктурированном тексте и превращать ее в структурированные данные, имеющие связи, которые могут быть проанализированы.

Вариант такой системы в настоящее время реализован и используется в качестве компоненты системы конкурентной разведки X-SCIF украинской компании «Информационная корпоративная служба», которая позволяет пользователю в онлайн-режиме получать карты связей для выбранных объектов и помогает интерпретировать результаты. Предусматривается, что пользователь вводит в качестве запроса объект. Запрос направляется к базе данных связей, откуда выбира-

ются соответствующие ему фрагменты – карты связей (уровень детализации и временная ретроспектива должны указываться параметрически).

После выявления релевантных объектов и связей выполняются процедуры их автоматической группировки (кластеризации) и визуализации, результаты предъявляются пользователю в виде карт связей, которые представляются в виде динамических (чаще всего, Java-диаграмм) графов связей.

В частности, в системе конкурентной разведки X-SCIF граф связей строится с помощью апплетов Java и представляет собой графический объект, который содержит в своем составе узлы и ребра. Каждый элемент графа связей имеет контекстное меню, которое является дополнительным элементом управления в интерфейсе пользователя (Рис. 11).

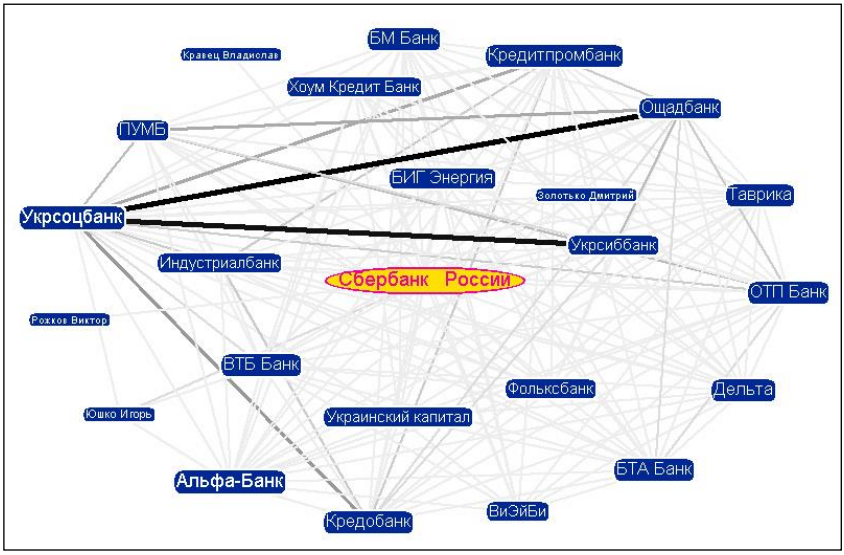


Рис. 11 – Граф информационных связей понятия «Сбербанк России»

Объекты, которые имеют большее количество связей, изображаются с помощью большего шрифта. Ребра, соответствующие большему количеству связей, изображаются более темными линиями. Построенная сеть имеет собственные средства управления: изменение масштаба (с помощью меню «масштаб» или полосы прокрутки в верхней части экрана);

перемещение всего графа; перемещение объекта; изменение конфигурации; подсветка связей выбранного узла и т.п.

На рис. 11 приведен пример использования базы данных связей, случай, когда пользователя интересуют информационные связи Сбербанка России по состоянию на 2011 год. Разумеется, для соответствующего запроса может быть выявлено множество различных связей, но при этом существует простой и надежный критерий ранжирования результатов, состоящий в отсечении статистического фона. В рассматриваемом случае, задав соответствующий запрос можно получить граф наиболее связанных со Сбербанком России объектов (персон и компаний). И если нахождение фамилий руководителей банка (председателя правления, первого заместителя председателя правления и руководителя дочернего банка) является достаточно очевидным результатом, то связи между отдельными банками позволили выявить (после обращения к документам-первоисточникам) неочевидные на первый взгляд факты, например, то, что УкрСиббанк и УкрСоцбанк являлись банками-партнерами.

Представленный подход может рассматриваться как основа построения так называемых «вертикальных» (предметно-ориентированных) информационно-поисковых систем, в которых изначально решены вопросы оперативности, отсеивания информационного шума. Рассматриваемая реализация имеет свойство масштабирования по трем параметрам: объему баз данных, составу понятий, которые используются, и по инфраструктурному окружению.

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

ставленный подход реализует связующее звено между полнотекстовыми и фактографическими базами данных.

2.5. Концепция Big Data

2.5.1. Понятие Больших Данных

Термин Big Data появился как новый термин и логотип в редакционной статье Клиффорда Линча, редактора журнала Nature 3 сентября 2008 года, который посвятил целый специальный выпуск одного из самых знаменитых журналов теме "что могут значить для современной науки наборы больших данных". В настоящее время этот термин уже прижился и достиг пика своего использования. Здесь слово "большие" было связано не столько с каким-то количеством, а с качественной оценкой. Время подтвердило справедливость выделения больших данных как отдельного феномена. Сегодня, согласно исследованиям агентства Gartner термин Big Data уже перешугнуло пик знаменитого гартнеровского Hype Cycle. На Рис. 12 приведена статистика запросив пользователей к системе Google по словосочетанию "Big Data" (сервис Google Trends, https://trends.google.com/).

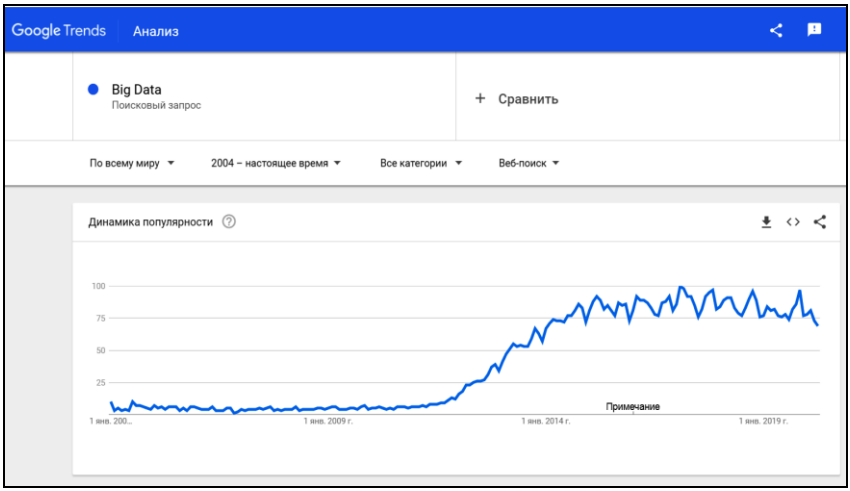


Рис. 12 – Динамика запросов "Big Data"

В 2012 году в статье [Boyd, 2012] Даны Бойд и Кэт Крауфорд было сформулировано определение Big Data как культурного, технологического и научного феномена, включающего в себя: 1) Технология: максимизация вычислительно мощности и сложности алгоритмов для сбора, анализа, связывания и сравнения огромных наборов данных. 2) Анализ: изображение огромных наборов данных чтобы идентифицировать паттерны для того, чтобы делать экономические, социальные технические и юридические утверждения. 3) Мифология: всеобщая уверенность, что огромные наборы данных представляют более высокую форму знаний и сведений, которые могут генерировать озарения, которые ранее были невозможны и с ореолом верности, объективности и точности.

Согласно этому определению, большие данные – это термин, обозначающий множество наборов данных столь объемных и сложных, что делает невозможным применение имеющихся традиционных инструментов управления базами данных и приложений для их обработки. Проблему представляют сбор, очистка, хранение, поиск, доступ, передача, анализ и визуализация таких наборов как целостной сущности, а не локальных фрагментов. В качестве определяющих характеристик для больших данных отмечают «три V»: объём (англ. volume, в смысле величины физического объёма), скорость (англ. Velocity, означающее в данном контексте скорость прироста и необходимость высокоскоростной обработки и получения результатов), многообразие (англ. variety, в смысле возможности одновременной обработки различных типов структурированных и полуструктурированных данных). Ведущей характеристикой здесь является объем данных, который должен быть рассмотрен в аспекте приложений.

Почему объем данных превратился в проблему? По мере того как компьютеры становились быстрее, а размер памяти больше, рос и объем данных. На самом деле, рост данных даже опережал рост быстродействия компьютеров, а лишь немногие алгоритмы линейно масштабируются с ростом входных данных. Короче говоря, данные растут быстрее, чем наша способность их обрабатывать. Таким образом, объем данных растет быстрее, чем обрабатывающие мощности. Отсюда вытекает ряд следствий.

- Некоторые методы и приемы, хорошо зарекомендовавшие себя в прошлом, теперь нуждаются в пересмотре или замене, потому что не масштабируются на современный объем данных.
- Алгоритмы не могут предполагать, что все исходные данные умещаются в оперативной памяти.
- Управление данными само по себе становится нетривиальной задачей.
- Применение кластеров или многоядерных процессоров становится необходимостью, а не роскошью.

Современность демонстрирует нам примеры чудовищных размеров генерируемых сегодня оцифрованных данных. Как утверждают гиганты ИТ индустрии (EMC, Cisco, IBM, Google) в 2012 году в мире было сгенерировано 2 зетабайта (2 * 1021) или 2 тысячи экзабайтов или 2 тысячи миллиардов гигабайтов информации, а в 2020 году эта величина достигнет 35 зетабайтов. Источниками этой лавины данных являются многочисленные цифровые устройства, концентрирующие и направляющие в бездонные просторы Интернета продукцию человеческого разума – твиты, посты в фэйсбуке и в контакте, запросы в поисковые системы, и т.п., а также данные от сенсоров и контроллеров миллионов устройств, которые измеряют температуру и влажность, состояние дорог и кондиционеров и много другого, что сегодня объединяется термином “Интернет вещей” IOT (Internet of Things). Однако этому препятствует не только проблема количества – объем данных первая “V”. Для больших данных, как было уже отмечено важна вторая “V”- скорость. Результаты обработки больших данных должны быть получены за время, определяемое решаемой с их помощью проблемы. Это даст возможность превратить аналитику больших данных из инструмента, отвечающего на вопрос “кто виноват?”, характерного для традиционных систем аналитики, в инструмент для получения ответов “что делать?”. Аналитик в этом случае из врача патологоанатома превращается в терапевта. Скорость доступа к данным, скорость их процессинга является важным критерием качества технологий, входящих в большие данные.

Наконец третья “V” – разнообразие данных говорит о том, что большие данные должны эффективно обрабатываться независимо от их структурированности. Здесь приня-

то выделять три основных вида данных по степени их структурированности.

Первый уровень – это привычные структурированные данные, которые могут быть представлены отделимыми и заранее определенными полями, в которых находятся биты, имеющие различную семантику. Например, все таблицы имеют в определенном поле заданной длины заголовки, в другом заранее заданном поле – один из фактов, в другом поле – другой из фактов, определяющих числовые или текстовые значения семантических переменных, содержащихся в заголовках.

Структурированные данные хорошо хранить в реляционных базах данных и управлять такими данными удобно, используя специальный язык SQL – Structured Query Language. Несмотря на свою распространенность такие данные определяют только в 10 % от всего объема сгенерированных данных.

Второй уровень – это полуструктурированные (semistructured) данные. Данные такого типа имеют структурные разделители, но не могут быть представлены в виде таблицы из-за отсутствия части атрибутов у разных данных. Примером таких данных могут служить файлы в формате SGML – Standard Generalized Markup Language или BibTex в которых нет определенной схемы хранения данных, но семантический смысл различных элементов данных может быть определен по анализу самого файла. Иногда такие данные определяют как допускающие самоописание. Многие данные хранящиеся в Web относятся к полуструктурированным, данные библиографических описаний публикаций, научные данные.

Наконец, неструктурированные данные, которые по определению не могут подойти под ранее описанные виды. В них входят тексты, записанные символами различных языков, записи звуков, неподвижные изображения, видеофайлы, сообщения электронной почты, твиты, презентации и другая бизнес-информация вне выгрузок баз данных. Считается, что от 80 до 90 процентов всех данных в организациях относятся к неструктурированным данным. Нередко к неструктурированным относят и введенные выше полуструктурированные данные. Иногда шкалу разнообразия расширяют,

используя целую шкалу от структурированных данных к полностью неструктурированным. Будем считать показатель вариативности данных нулевым для полностью неструктурированных данных и возрастающим до единицы для хорошо структурированных из реляционных баз. По мнению участников Всемирного экономического форума 2012 года в Давосе, те, кто оседлает тему интеллектуального анализа больших данных, станут хозяевами информационного пространства. Этой теме был посвящен специальный доклад на Форуме «Большие данные – большое влияние». Ключевой вывод доклада – цифровые активы становятся не менее значимым экономическим активом, чем золото или валюта. Исследования, проведенные профессором Бринйолфсоном (E.Brynjolfsson) и двумя его коллегами в 2012 году, показали, что анализ и прогнозирование на основе больших данных берется на вооружение корпоративной Америкой. Они изучили 179 крупных компаний и обнаружили, что те из них, кто взял в последние год- полтора на вооружение интеллектуальный анализ больших данных получил немедленное улучшение экономических показателей на 5-6%. В настоящее время потребности общества делают необходимым появление специалистов по большим данным в форме отдельной профессии. Название этой профессии Data Scientist – исследователь данных. В США произвели оценку потребностей в специалистах такой профессии и пришли к выводу, что уже в 2018 году в США будет нехватка исследователей данных в количестве 190 000 человек! Известный журнал Harvard Business Review так озаглавил один из своих выпусков: “Data Scientist: The Sexiest Job Of the 21st Century” – исследователь данных – самая привлекательная работа 21 столетия.

2.5.2. Техники больших данных

Сначала перечислим те функциональные операции над данными, методы их хранения и обработки. Разумеется, этот список не исчерпывает всего многообразия динамично развивающихся техник, однако позволяет увидеть, что можно делать с большими данными для достижения целей, стоящих перед исследователем.

- Консолидация данных;
- Классификация, кластеризація;

- Машинное обучение;
- Визуализация.

Консолидация данных

Этот целый набор техник, направленных на извлечение данных из разных источников, обеспечение их качества, преобразования в единый формат и загрузку в хранилище данных – “аналитическую песочницу” (analytic sandbox) или «озеро данных» (data lake). Техники консолидации данных различаются по виду аналитики выполняемой системой :

- Пакетная аналитика (batch oriented);
- Аналитика реального времени (real time oriented);
- Гибридная аналитика (hybrid).

При пакетной аналитике периодически производится выгрузка данных из различных источников, данные анализируются на наличие сбойных фрагментов, шума и производится их фильтрация. При выполнении аналитики реального времени данные производятся источниками непрерывно и образуют набор потоков данных. Анализ этих потоков и своевременное получение результатов в заданном темпе требуют обеспечить асинхронное получение данных в виде некоторых сообщений и маршрутизировать эти сообщения в нужные процессинговые узлы для обработки. Для гибридной аналитики как правило сообщения данных должны быть не только маршрутизированы на процессинг, но и интегрированы в аналитическую песочницу для дальнейшей обработки по результатам накопления данных за значительные интервалы времени. Данные, полученные в результате консолидации, должны соответствовать определенным критериям качества. Качество данных - это критерий, определяющий полноту, точность, актуальность и возможность интерпретации данных. Данные могут быть высокого и низкого качества. Данные высокого качества - это полные, точные, актуальные данные, которые поддаются интерпретации. Такие данные обеспечивают получение качественного результата: знаний, которые смогут поддерживать процесс принятия решений. овокупность процессов, определяющих консолидацию, называют ETL – Extraction-Transformation-Loading (Извлечение-Преобразование-Загрузка). В приложения бизнес-аналитики в процессы ETL включались весьма сложные преобразования

данных, такие как квантование, позволяющее снизить объем обрабатываемых данных, нормализация – процесс приведения реляционных таблиц к каноническому виду или числовых данных к единому масштабу, кодирование данных – введение уникальных кодов для сжатия данных. В техниках больших данных обычно полагают, что необходимо работать непосредственно с грязными данными, поскольку нередко именно характер сбоев может стать предметом анализа, а сжатие данных представляет собой функцию собственно аналитических алгоритмов. Возможность же хранения данных в исходном виде должны предоставлять технические средства аналитической системы. Качество больших данных нередко трудно оценить методами формальных алгоритмов, и тогда прибегают к визуализации на раннем этапе исследования. Кроме оценки качества и выбора метода препроцессинга, визуализация может помочь перейти к важному этапу аналитики - выбору моделей, гипотез для достижения конечной цели - принятия решений.

Визуализация

Техника визуализации является мощным методом интеллектуального анализа данных. Как правило, ее используют для просмотра и верификации данных перед созданием модели, а также после генерации прогнозов. Визуализация - это преобразование численных данных в некоторый визуальный образ, в целях упрощения восприятия больших массивов информации. Для осуществления визуализации служат визуализаторы. Визуализаторы могут являться либо отдельным приложением, либо плагином или частью другого приложения. Возможности визуализаторов очень широки. В настоящее время они могут представлять информацию практически во всех мыслимых видах, лишь бы аналитик мог сформулировать, что он хочет видеть.

Визуализация текстов

Если данные представляют собой тексты на естественном языке, то первичную помощь в анализе может оказать визуализация с помощью размеченного текста. Визуализатор подсчитывает частоту упоминаний того или иного слова, и присваивает словам условный вес, зависящий от этой часто-

ты. Слова разного веса при визуализации имеют различную разметку, а значит разное представление на экране. Одни слова выглядят больше других. Этот тип визуализации помогает исследователю очень быстро ухватить основные мысли текста.

Визуализация кластеров

Одной из часто используемых визуализаций является визуализация кластеров. Кластерами называют группы в чем-то схожих или близких по свойствам объектов. Алгоритмы кластеризации, т.е. разбиение множества объектов на группы, мы рассмотрим ниже, а здесь покажем только как может быть визуализирована их работа. Большинство визуализаторов поддерживает алгоритмы кластеризации и способно разделять данные на кластеры. Обычно для визуального представления кластеров для объектов из разных кластеров используются контрастные цвета.

Визуализация ассоциаций

Визуализация ассоциаций демонстрирует частоту, с которой те или иные элементы появляются вместе в наборе данных, за счет чего определяется структура организации данных (например, речь может идти о том, какие продукты часто продаются вместе). Также возможна визуализация информации о силе ассоциации данных.

Визуализация гипотез

Визуализация гипотез позволяет показывать выявленные закономерности, подтверждающие выдвигаемые гипотезы. Представление информации в различных визуализаторах отличается. Например, если строки круговых 3D-диаграмм отображают признаки, использованные классификатором, то каждая круговая диаграмма отражает вероятность того, что величина признака или диапазон значений подходят для классификации. На рисунке 2.10, представленном ниже, анализируется зарплата работающего населения США. Визуализатор отражает атрибуты, которые могут влиять на классификацию по зарплате. Атрибуты представлены рядами круговых трехмерных диаграмм. Высота круговой диаграммы (цилиндра) показывает количество записей в данной ка-

тегории; цвет показывает, что зарплата больше или меньше 50 тыс. долл. На каждый атрибут может быть несколько круговых диаграмм, например, для обозначения пола (мужской/женский) имеется две диаграммы, а для возраста — восемь диаграмм. Их количество зависит от количества закономерностей, выявленных визуализатором.

Визуализация деревьев решений

Визуализация деревьев решений позволяет представить иерархически организованную информацию в виде ландшафта и обозревать все множество данных или их часть в виде узлов и ветвей. Ландшафт может быть как двумерным, так и трехмерным. Количественные и реляционные характеристики данных делаются видимыми с помощью иерархически соединенных узлов.

Классификация

Техника классификации является одной из базовых методик интеллектуального анализа больших данных. Ее нередко используют при построении модели аналитических систем наряду с еще одной техникой - кластеризацией. Классификация - это распределение объектов (наблюдений, событий) исследования по заранее известным классам на основании сходства признаков. В отличие от классификации кластеризация производит распределение объектов (наблюдений, событий) по неизвестным заранее классам. Классификация производится в соответствии с принципами машинного обучения с учителем (Supervised Machine Learning). Для проведения классификации с помощью математических методов необходимо иметь формальное описание объекта, которым можно оперировать, используя математический аппарат классификации. Каждый объект (запись базы данных) должен содержать информацию о некоторых признаках объекта.

Процесс классификации, как правило, сводится к следующим шагам.

1. Набор исходных данных (или выборку данных) разбивают на два множества: обучающее и тестовое. Обучающее множество - множество, которое включает данные, использующиеся для конструирования модели. Множество содержит входные и выходные (целевые) значения примеров. Вы-

ходные значения предназначены для обучения модели. Тестовое множество также содержит входные и выходные значения примеров. Здесь выходные значения используются для проверки модели.

2. Каждый объект набора данных относится к одному предопределенному классу. На этом этапе используется обучающее множество, на нем происходит конструирование модели. Полученная модель представляется классификационными правилами, деревом решений или математическими формулами.

3. Производится оценка правильности модели. Известные значения из тестового множества сравниваются с результатами использования полученной модели. Вычисляется уровень точности - процент правильно классифицированных объектов в тестовом множестве.

Кластеризация

Техника кластеризации является подходом к классификации данных в случае, когда заранее неизвестно, к какому классу должен быть отнесен любой из имеющихся объектов. Кластеризация осуществляется автоматическим нахождением групп, на которые должны быть разбиты анализируемые объекты. Такой процесс может рассматриваться как машинное обучение без учителя (Unsupervised Machine Learning). Известно более 100 разных алгоритмов

Машинное обучение

Термин «машинное обучение», скорее всего, встречался вам не раз. Хотя его нередко используют как синоним искусственного интеллекта, на самом деле машинное обучение – это один из его элементов. При этом оба понятия родились в Массачусетском технологическом институте в конце 1950-х годов.

Машинное обучение (machine learning, ML) — класс методов искусственного интеллекта, характерной чертой которых является не прямое решение задачи, а обучение в процессе применения решений множества сходных задач. Для построения таких методов используются средства математической статистики, численных методов, методов оптимизации, теории вероятностей, теории графов, различные техники работы с данными в цифровой форме.

Различают два типа обучения:

Обучение по прецедентам, или индуктивное обучение, основано на выявлении эмпирических закономерностей в данных.

Дедуктивное обучение предполагает формализацию знаний экспертов и их перенос в компьютер в виде базы знаний.

Дедуктивное обучение принято относить к области экспертных систем, поэтому термины машинное обучение и обучение по прецедентам можно считать синонимами.

Многие методы индуктивного обучения разрабатывались как альтернатива классическим статистическим подходам. Многие методы тесно связаны с извлечением информации (information extraction, information retrieval), интеллектуальным анализом данных (data mining).

В отличие от традиционного ПО, которое прекрасно справляется с выполнением инструкций, но не способно к импровизации, системы машинного обучения по сути программируют сами себя, самостоятельно разрабатывая инструкции путем обобщения известных сведений.

Классический пример – распознавание образов. Покажите системе машинного обучения достаточное количество снимков собак с пометкой «собака», а также кошек, деревьев и других объектов, помеченных «не собака», и она со временем начнет хорошо отличать собак. И для этого ей не нужно будет объяснять, как именно те выглядят.

Обучение с учителем и без

Упомянутый вид машинного обучения называется обучением с учителем. Это значит, что кто-то познакомил алгоритм с огромным объемом учебных данных, просматривая результаты и корректируя настройки до тех пор, пока не была достигнута нужная точность классификации данных, которые система еще не «видела». Это то же самое, что нажимать кнопку «не спам» в почтовой программе, когда фильтр случайно перехватывает нужное вам сообщение. Чем чаще вы это делаете, тем точнее становится фильтр.

Типичные задачи обучения с учителем – классификация и прогнозирование (или регрессионный анализ). Распознава-

ние спама и образов – задачи классификации, а прогнозирование котировок акций – классический пример регрессии.

При обучении без учителя система просматривает гигантские объемы данных, запоминая, как выглядят «нормальные» данные, чтобы получить возможность распознавать аномалии и скрытые закономерности. Обучение без учителя полезно, когда вы точно не знаете, что именно ищете, – в этом случае систему можно заставить вам помочь.

Системы обучения без учителя могут обнаруживать закономерности в огромных объемах данных гораздо быстрее, чем люди. Именно поэтому банки используют их для выявления мошеннических операций, маркетологи – для идентификации клиентов со схожими атрибутами, а ПО безопасности – для распознавания вредоносной активности в сети.

Примеры задач обучения без учителя – кластеризация и поиск правил ассоциации. Первая применяется, в частности, для сегментации клиентов, а на поиске правил ассоциации основаны механизмы выдачи рекомендаций.

Способы машинного обучения

Раздел машинного обучения, с одной стороны, образовался в результате разделения науки о нейронных сетях на методы обучения сетей и виды топологий их архитектуры, с другой стороны — вобрал в себя методы математической статистики. Указанные ниже способы машинного обучения исходят из случая использования нейросетей, хотя существуют и другие методы, использующие понятие обучающей выборки — например, дискриминантный анализ, оперирующий обобщённой дисперсией и ковариацией наблюдаемой статистики, или байесовские классификаторы. Базовые виды нейросетей, такие как перцептрон и многослойный перцептрон (а также их модификации), могут обучаться как с учителем, так и без учителя, с подкреплением и самоорганизацией. Но некоторые нейросети и большинство статистических методов можно отнести только к одному из способов обучения. Поэтому, если нужно классифицировать методы машинного обучения в зависимости от способа обучения, будет некорректным относить нейросети к определенному виду, правильнее было бы типизировать алгоритмы обучения нейронных сетей.

- Обучение с учителем — для каждого прецедента задаётся пара «ситуация, требуемое решение».
- Обучение без учителя — для каждого прецедента задаётся только «ситуация», требуется сгруппировать объекты в кластеры, используя данные о попарном сходстве объектов, и/или понизить размерность данных.
- Активное обучение — отличается тем, что обучаемый алгоритм имеет возможность самостоятельно назначать следующую исследуемую ситуацию, на которой станет известен верный ответ.
- Обучение с частичным привлечением учителя (semi-supervised learning) — для части прецедентов задается пара «ситуация, требуемое решение», а для части — только «ситуация».
- Трансдуктивное обучение — обучение с частичным привлечением учителя, когда прогноз предполагается делать только для прецедентов из тестовой выборки.
- Многозадачное обучение (multi-task learning) — одновременное обучение группе взаимосвязанных задач, для каждой из которых задаются свои пары «ситуация, требуемое решение».
- Многовариантное обучение (multiple-instance learning) – обучение, когда прецеденты могут быть объединены в группы, в каждой из которых для всех прецедентов имеется «ситуация», но только для одного из них (причем, неизвестно какого) имеется пара «ситуация, требуемое решение»
- Бустинг (boosting — улучшение) — это процедура последовательного построения композиции алгоритмов машинного обучения, когда каждый следующий алгоритм стремится компенсировать недостатки композиции всех предыдущих алгоритмов.
- Байесовская сеть.

Алгоритмы машинного обучения нуждаются в данных, в как можно большем количестве данных из как можно более широкого набора источников. Чем больше они "питаются" этими данными, тем "умнее" становятся и тем больше их потенциал при принятии решений. И облака дают эти большие данные.

Большие данные сулят нам найти много ценного в процессе цифровой трансформации, в то время как облако предлагает строительные блоки для этого процесса. Машинное обучение, в свою очередь, стало первым по-настоящему промышленным инструментом для масштабного освоения этих новых ценностей. Привлекательность машинного обучения в том, что возможности его использования практически безграничны. Оно может применяться везде, где важен быстрый анализ данных, и оказать просто-таки революционный эффект там, где важно выявлять тенденции или аномалии в обширных наборах данных — от клинических исследований до сферы безопасности и контроля за соблюдением стандартов.

Ограничения машинного обучения

Каждая система машинного обучения создает собственную схему связей, представляя собой нечто вроде черного ящика. Вы не сможете путем инженерного анализа выяснить, как именно выполняется классификация, но это и не имеет значения, главное, чтобы работало.

Однако система машинного обучения хороша лишь настолько, насколько точны учебные данные: если подать ей на вход «мусор», то и результат будет соответствующим. При неправильном обучении или слишком малом размере обучающей выборки алгоритм может выдавать неверные результаты.

2.5.3. Технологии и инструменты больших данных

Мы рассмотрим базовые технологии и инструменты, которые сегодня получили наибольшее распространение в известных проектах. Этот список не исчерпывает всех уже апробированных технологий и тем более находящихся в разработке, однако он позволяет получить достаточно целостное представление о том “чем” пользуются сегодня исследователи данных и какими инструментами необходимо владеть, чтобы развернуть проект с использованием больших данных.

Технологии больших данных должны обеспечивать решениями и инструментами, позволяющими реализовывать описанные выше техники на значительных объемах разнородных данных с необходимой скоростью. Достигается это вы-

сокой параллелизацией вычислений и распределенным хранением данных. Несмотря на потребность значительной вычислительной мощности и памяти, как правило, развертывание программных продуктов больших данных производится на кластерах из компьютеров среднего или даже низкого класса (commodity computers). Это позволяет масштабировать системы больших данных без привлечения существенных затрат. В последнее время для развертывания систем больших данных все шире применяются облачные сервисы (cloud computing services). В случае имплементации системы в облаке узлы вычислительного кластера реализуются на виртуальных машинах облачной инфраструктуры и гибко адаптируются к задаче, снижая затраты на использование. Это служит дополнительным фактором, привлекающим многих разработчиков строить системы больших данных на облачных платформах.

Наиболее популярной технологией больших данных, считающейся де-факто стандартом для построения систем аналитики, работающих в пакетном режиме, является совокупность решений и программных библиотек, объединенных под названием Hadoop. Если большие данные поступают в виде высокоскоростных потоков и реагирование системы должно происходить с малой задержкой, то вместо пакетной аналитики применяется аналитика реального времени. Здесь пока не возникло де-факто стандартных подходов и из наиболее популярных мы рассмотрим технологию под названием Storm.

Apache Hadoop

Под названием Hadoop сообщество Apache продвигает технологию, основанную на использовании специальной инфраструктуры для параллельной обработки больших объемов данных. Hadoop обеспечивает среду для функционального программирования задач, автоматического распараллеливания работ, смещения вычислительной нагрузки к данным. Hadoop создал Дуг Каттинг — создатель Apache Lucene, широко используемой библиотеки текстового поиска. Hadoop происходит от Apache Nutch — системы вебпоиска с открытым кодом, которая сама по себе являлась частью проекта Lucene.

Проект Nutch был запущен в 2002 году. Работоспособный обходчик и поисковая система появились очень быстро. Однако разработчики поняли, что их архитектура не будет масштабироваться на миллиарды веб-страниц. Помощь пришла в 2003 году, когда была опубликована статья с описанием архитектуры GFS (Google File System) — распределенной файловой системы, которая использовалась в реальных проектах Google2 [Ghemawat, 2003].

В 2004 году была опубликована статья, в которой компания Google представила миру технологию MapReduce [Dean, 2004]. В начале 2005 года у разработчиков Nutch появилась работоспособная реализация MapReduce на базе Nutch, а к середине года все основные алгоритмы Nutch были адаптированы для использования MapReduce и NDFS. Возможности применения NDFS и реализации MapReduce в Nutch выходили далеко за рамки поиска, и в феврале 2006 года был образован независимый подпроект Lucene, получивший название Hadoop. Примерно в то же время Дуг Каттинг поступил в компанию Yahoo!, которая предоставила группу и ресурсы для превращения Hadoop в систему, работающую в веб-масштабах (см. далее врезку «Hadoop в Yahoo!»). Результаты были продемонстрированы в феврале 2008 года, когда компания Yahoo! объявила, что используемый ею поисковый индекс был сгенерирован 10000-ядерным кластером Hadoop [Yahoo, 2008].

История Hadoop непосредственно связана с разработкой Google File System (2003 год) и затем реализацией технологии MapReduce (2004 год). На основе этих компонент в 2005 году появилось приложение поиска информации Apache Nutch, которое на следующий год дало дорогу проекту Apache Hadoop.

Хотя Hadoop чаще всего ассоциируется с MapReduce и распределенной файловой системой (HDFS, ранее называвшейся NDFS), этим термином часто обозначают целое семейство взаимосвязанных проектов, объединенных инфраструктурой распределенных вычислений и крупномасштабной обработки данных. Все базовые проекты, рассматриваемые в книге, ведутся фондом Apache Software Foundation, предоставляющим поддержку сообщества проектов с открытым кодом — включая исходный HTTP-сервер, от которого произошло

название. С расширением экосистемы Hadoop появляются новые проекты, не обязательно находящиеся под управлением Apache, но предоставляющие дополнительные функции Hadoop или образующие абстракции более высокого уровня на основе базовой функциональности. 42 Глава 1. Знакомство с Hadoop Ниже кратко перечислены проекты Hadoop, рассмотренные в книге. Common — набор компонентов и интерфейсов для распределенных файловых систем и общего ввода/вывода (сериализация, Java RPC, структуры данных). Avro — система сериализации для выполнениях эффективных межъязыковых вызовов RPC и долгосрочного хранения данных. MapReduce — модель распределенной обработки данных и исполнительная среда, работающая на больших кластерах типовых машин. HDFS — распределенная файловая система, работающая на больших кластерах стандартных машин. Pig — язык управления потоком данных и исполнительная среда для анализа очень больших наборов данных. Pig работает в HDFS и кластерах MapReduce. Hive — распределенное хранилище данных. Hive управляет данными, хранимыми в HDFS, и предоставляет язык запросов на базе SQL (которые преобразуются ядром времени выполнения в задания MapReduce) для работы с этим данными. HBase — распределенная столбцово-ориентированная база данных. HBase использует HDFS для организации хранения данных и поддерживает как пакетные вычисления с использованием MapReduce, так и точечные запросы (произвольное чтение данных). ZooKeeper — распределенный координационный сервис высокой доступности. ZooKeeper предоставляет примитивы, которые могут использоваться для построения распределенных приложений (например, распределенные блокировки). Sqoop — инструмент эффективной массовой пересылки данных между структурированными хранилищами (такими, как реляционные базы данных) и HDFS. Oozie — сервис запуска и планирования заданий Hadoop (включая задания MapReduce, Pig, Hive и Sqoop jobs)

Hadoop состоит из четырех функциональных частей:

- Hadoop Common;
- Hadoop HDFS;
- Hadoop MapReduce;
- Hadoop YARN.

Hadoop Common – это набор библиотек и утилит, необходимых для нормального функционирования технологии. В его состав входит специализированный упрощённый интерпретатор командной строки.

Когда набор данных перерастает емкость одной физической машины, его приходится распределять по нескольким разным машинам. Файловые системы, управляющие хранением данных в сети, называются распределенными файловыми системами. Поскольку они работают в сетевой среде, проектировщику приходится учитывать все сложности сетевого программирования, поэтому распределенные файловые системы сложнее обычных дисковых файловых систем. Например, одна из самых серьезных проблем — сделать так, чтобы файловая система переживала сбои отдельных узлов без потери данных. Hadoop поставляется с распределенной файловой системой, которая называется HDFS (Hadoop Distributed Filesystem). Иногда — в старой документации или конфигурациях или в неформальном общении — также встречается сокращение «DFS»; оно означает то же самое. HDFS — основная файловая система Hadoop, которой посвящена эта глава, но в Hadoop также реализована абстракция обобщенной файловой системы, и мы попутно рассмотрим интеграцию Hadoop с другими системами хранения данных (например, локальной файловой системой и Amazon S3).

HDFS (Hadoop Distributed File System) – это распределенная файловая система для хранения данных на множестве машин в больших объемах. Проектировалась так, чтобы обеспечивать:

- Надежное хранение данных на дешевом
- ненадежном оборудовании;
- Высокую пропускную способность чтения-записи;
- Потоковый доступ к данным;
- Упрощенную модель согласованности;
- Архитектуру аналогичную Google File System.

Файловая система HDFS спроектирована для хранения очень больших файлов с потоковой схемой доступа к данным в кластерах обычных машин [Shvachko, 2010]. Рассмотрим это утверждение более подробно. Очень большие файлы Под «очень большими» в этом контексте подразумеваются файлы, размер которых составляет сотни мегабайт, гигабайт и тера-

байт. Сейчас существуют кластеры Hadoop, в которых хранятся петабайты даннях [Scaling, 2008].

Потоковый доступ к данным В основу HDFS заложена концепция однократной записи/многократного чтения как самая эффективная схема обработки данных. Набор данных обычно генерируется или копируется из источника, после чего с ним выполняются различные аналитические операции. В каждой операции задействуется большая часть набора данных (или весь набор), поэтому время чтения всего набора данных важнее задержки чтения первой записи. Обычное оборудование Hadoop не требует дорогостоящего оборудования высокой надежности. Система спроектирована для работы на стандартном оборудовании (общедоступное оборудование, которое может быть приобретено у многих фирм)3 с достаточно высокой вероятностью отказа отдельных узлов в кластере (по крайней мере, для больших кластеров).

Технология HDFS спроектирована таким образом, чтобы в случае отказа система продолжала работу без сколько-нибудь заметного прерывания. Также следует выделить области применения, для которых в настоящее время HDFS подходит не лучшим образом (при том, что в будущем ситуация может измениться): Быстрый доступ к данным Приложения, требующие доступа к данным с минимальной задержкой (в диапазоне десятков миллисекунд), плохо сочетаются с HDFS. Напомним, что система HDFS оптимизирована для обеспечения высокой пропускной способности передачи данных, за которую приходится расплачиваться замедлением доступа. HBase (глава 13) в настоящее время лучше подходит для организации доступа к данным с минимальной задержкой.

Многочисленные мелкие файлы Так как узел имен хранит метаданные файловой системы в памяти, предел количества файлов в файловой системе определяется объемом памяти узла имен. Как показывает опыт, каждый файл, каталог и блок занимают около 150 байт. Таким образом, например, если у вас имеется миллион файлов, каждый из которых занимает один блок, для хранения информации потребуется не менее 300 Мбайт памяти. Хранение миллионов файлов еще приемлемо, но миллиарды файлов уже выходят за пределы возможностей современного оборудования1 . Множественные источники записи, произвольные модификации файлов За-

пись в файлы HDFS может выполняться только одним источником. Запись всегда осуществляется в конец файла. Поддержка множественных источников записи или модификации с произвольным смещением в файле отсутствует. (Может быть, эти возможности будут поддерживаться в будущем, но, скорее всего, они будут относительно неэффективными).

В основе архитектуры HDFS лежат узлы хранения – серверы стандартной архитектуры, на внутренних дисках которых хранятся данные. Для всех данных используется единое адресное пространство. При этом обеспечивается параллельный ввод-вывод информации с разных узлов. Таким образом, гарантируется высокая пропускная способность системы.

HDFS оперирует на двух уровнях: пространства имён (Namespace) и хранения блоков данных (Block Storage Service). Пространство имён поддерживается центральным узлом имён (Namenode), хранящим метаданные файловой системы и метаинформацию о распределении блоков файлов.

Многочисленные узлы данных (Datanode) непосредственно хранят файлы. Узел имён отвечает за обработку операций файловой системы— открытие и закрытие файлов, манипуляция с каталогами и т.п. Узлы данных отрабатывают операции по записи и чтению данных. Узел имён и узлы данных снабжаются веб-серверами, отображающими текущий статус и позволяющими просматривать содержимое файловой системы.

У HDFS нет POSIX-совместимости. Не работают Unix-команды ls, cp и т.п. Для монтирования HDFS в Linux ОС необходимы специальные инструменты, например, HDFS-Fuse. Файлы поблочно распределяются между узлами. Все блоки в HDFS (кроме последнего блока файла) имеют одинаковый размер – от 64 до 256 Мб.

Для обеспечения устойчивости к отказам серверов, каждый блок может быть продублирован на нескольких узлах. Коэффициент репликации (количество узлов, на которых должен быть размещён каждый блок) определяется в настройках файла. Файлы в HDFS могут быть записаны лишь однажды (модификация не поддерживается), а запись в файл в одно время может вести только один процесс. Таким простым образом реализуется согласованность данных.

Hadoop MapReduce

MapReduce — модель программирования, ориентированная на обработку данных. Эта модель проста, но не настолько, чтобы в ее контексте нельзя было реализовать полезные программы. Hadoop позволяет запускать программы MapReduce, написанные на разных языках; в этой главе мы рассмотрим одну и ту же программу, написанную на языках Java, Ruby, Python и C++. Но самое важное заключается в том, что программы MapReduce параллельны по своей природе, а следовательно, крупномасштабный анализ данных становится доступным для всех, у кого в распоряжении имеется достаточно компьютеров. Достоинства MapReduce в полной мере проявляются в работе с большими наборами данных, так что начнем с рассмотрения одного из таких наборов.

Hadoop MapReduce – это наиболее популярная программная реализация модели параллельной обработки больших объемов данных путем разделения на независимые задачи, решаемые функциями Map и Reduce. Алгоритм MapReduce получает на вход 3 аргумента: исходную коллекцию данных, Map функцию, Reduce функцию, и возвращает результирующую коллекцию данных.

Исходными коллекциями данных являются наборы записей специального вида, Это структура данных типа Ключ,Значение (KEY, VALUE). Пользователю необходимо задать функции обработки Map и Reduce. Алгоритм сам заботится о сортировке данных, запуске функций обработки, повторном исполнении упавших транзакций и много чем еще. Результирующая коллекция состоит из результатов анализа в легко интерпретируемом виде. Работа алгоритма MapReduce состоит из трех основных этапов: Map, Group и Reduce. В качестве первого этапа над каждым элементом исходной коллекции выполняется Map функция. Как правило, она принимает на вход одну запись вида (KEY, VALUE), и возвращает по ней некоторое количество новых записей (KEY1, VALUE1), (KEY2, VALUE2), …, т.е. преобразует входную пару {ключ: значение} в набор промежуточных пар. Также эта функция играет роль фильтра — если для данной пары никаких промежуточных значений возвращать не нужно, функция возвращает пустой список.

Можно сказать, что обязанность Map функции конвертировать элементы исходной коллекции в ноль или несколько экземпляров объектов {ключ: значение}.

На втором этапе (Group) алгоритм сортирует все пары {ключ: значение}и создает новые экземпляры объектов, сгруппированные по ключу. Операция группирования выполняется внутри алгоритма MapReduce и пользователем не задается. Функция Reduce возвращает экземпляры объекта {ключ: свернутое значение}, которые включается в результирующую коллекцию.

Для примера, рассмотрим упрощенный вариант задачи, стоящей перед поисковыми системами. Допустим, у нас есть база данных страниц в Интернете, и мы хотим, сколько раз ссылаются на каждую страницу. Пусть есть страница first.com со ссылками на first.com, second.com, third.com, страница second.com с двумя ссылками на first.com и страница third.com, на которой нет ссылок вообще.

Чтобы иметь единый формат исходной коллекции данных, определим вид каждой сохраненной страницы как (KEY = URL, VALUE = TEXT). Результаты легко интерпретируются.

В качестве базового языка написания функций используется Java. Для программирования существует популярный Hadoop плагин в Eclipse. Но можно обойтись и без него: утилиты Hadoop streaming позволяют использовать в качестве Map и Reduce любой исполняемый файл, работающий со стандартным вводом-выводом операционной системы (например, утилиты командной оболочки UNIX, скрипты Python, Ruby и т.д.), есть также SWIG-совместимый прикладной интерфейс программирования Hadoop pipes на C++. Кроме того, в состав дистрибутивов Hadoop входят реализации различных обработчиков, наиболее часто используемых в распределённой обработке.

Особенностью Hadoop является перемещение вычислений как можно ближе к данным. Поэтому пользовательские задачи запускаются на том узле, который содержит данные для обработки. По окончанию фазы Map происходит перемещение промежуточных списков данных для обработки функцией Reduce. Заметим здесь, что кроме Hadoop существуют разные имплементации MapReduce. Изначально MapReduce был реализован компанией Google. Позднее появились другие

реализации алгоритма. Развитием MapReduce от Google стал проект с открытым исходным кодом - MySpace Qizmt - MySpace's Open Source Mapreduce Framework. Другой известной версией алгоритма является та, что реализована в системе MongoDB

Hadoop YARN (Yet Another Resource Negotiator) – платформа управления ресурсами системы, ответственная за распределение вычислительных ресурсов серверов и расписание выполнения пользовательских задач.

В первых версиях Hadoop MapReduce включал планировщик заданий JobTracker, начиная с версии 2.0 (2013 г.) эта функция перенесена в YARN. В ней модуль Hadoop MapReduce реализован поверх YARN. Программные интерфейсы по большей части сохранены, однако полной обратной совместимости нет.

YARN иногда называют кластерной операционной системой. Это обусловлено тем, что платформа ведает интерфейсом между аппаратными ресурсами и различными приложениями, использующими вычислительные мощности. Основой YARN является логически самостоятельный демон — планировщик ресурсов (ResourceManager), абстрагирующий все вычислительные ресурсы кластера и управляющий их предоставлением приложениям распределённой обработки. Ему подотчетны многочисленные менеджеры узлов (Node Manager), ответственные за отслеживание текущего статуса и нагрузки отдельных серверов.

Работать под управлением YARN могут как MapReduce-программы, так и любые другие распределённые приложения, поддерживающие соответствующие программные интерфейсы. YARN обеспечивает возможность параллельного выполнения нескольких различных задач в рамках системы серверов.

Разработчику распределённого приложения необходимо реализовать специальный класс управления приложением (AppMaster), который отвечает за координацию заданий в рамках тех ресурсов, которые предоставит планировщик ресурсов. Планировщик ресурсов отвечает за создание экземпляров класса управления приложением и взаимодействия с ними через сетевой протокол.

На основе Hadoop создан целый ряд продуктов для обработки данных. Вот список лишь наиболее популярных из них:

- Pig – высокоуровневый язык потоков данных для параллельного программирования;
- HBase – распределенная база данных, которая обеспечивает хранение больших таблиц;
- Cassandra – устойчивая к ошибкам, децентрализованная база данных;
- Hive – хранилище данных с функциями объединения данных и быстрого поиска;
- Mahout – библиотека методов машинного обучения и извлечения знаний.

Hadoop является очень динамично развивающейся технологией. Поэтому наиболее свежую информацию рекомендуется получать в Интернете на сайте http://hadoop.apache.org/ .

Storm – система потоковой обработки

Storm является бесплатной технологией и программной реализацией распределенной вычислительной системы реального времени [15]. Эта система позволяет строить надежную обработку неограниченных потоков данных подобно тому как Hadoop делает это с пакетной обработкой. Storm применяется для аналитики реального времени, онлайнового машинного обучения, непрерывных вычислений, распределенных ETL и других операций с потоками больших данных.

Storm может интегрироваться с технологиями очередей и баз данных, которые уже используются и не зависит от языка программирования. Основой Storm являются Storm топологии и Storm кластер. Кластер является объектом, подобным Hadoop кластеру, а вместо запуска MapReduce job здесь запускаются Storm topologies. Jobs и Topologies имеют ключевое различие – первые в нормальном режиме завершают работу, а вторые обрабатывают сообщения всегда. В Storm кластере имеется два типа узлов master node и worker nodes (рисунок 2.28). На master node запускается демон называемый Nimbus, который подобен JobTracker в Hadoop. Nimbus ответственнен за распределение кода по рабочим узлам клас-

тера, распределение задач по машинам и запуск и остановку рабочих процессов. Каждый рабочий процесс выполняет подмножество топологии. Работающая топология состоит из многих рабочих процессов, распределенных по многим машинам. Каждый рабочий узел (worker node) имеет демон под названием Supervisor. Этот модуль слушает все процессы на своей машине и запускает и останавливает их по инициативе Nimbus. Координация между Nimbus и всеми Supervisor производится через специальный кластер, называемый Zookeeper. Этот кластер также хранит на своем дисковом пространстве состояние всех процессов, что позволяет восстанавливать после сбоя отдельно любую машину рабочего кластера. Чтобы выполнить вычисления в реальном времени на Storm нужно создать топологию (topologies) – граф вычислений. Каждый узел в топологии содержит логику процессинга и линк между узлами, показывающий как данные должны быть переданы между узлами.

Основной абстракцией в Storm является поток (stream). Потоком называется неограниченная последовательность кортежей (tuples). Источники потоков данных для обработки представляются в топологии абстракцией, называемой spout, а обработчики потоков, которые могут выполнять функции, фильтровать потоки, агрегировать или объединять потоки данных, взаимодействовать с базами данных называются bolt.

Стек Elastic

За последние несколько лет появились различные системы для хранения и обработки больших массивов данных. Среди них можно выделить проекты экосистемы Hadoop, некоторые базы данных (БД) NoSQL, а также поисковые и аналитические системы наподобие Elasticsearch. Hadoop и любая база данных NoSQL имеют свои преимущества и области применения.

Elastic Stack — обширная экосистема компонентов, которые служат для поиска и обработки данных. Основные компоненты Elastic Stack — это Kibana, Logstash, Beats, X-Pack и Elasticsearch. Ядром Elastic Stack выступает поисковая система Elasticsearch, которая предоставляет возможности для хранения, поиска и обработки данных. Утилита

Kibana, которую также называют окном в Elastic Stack, является отличным средством визуализации и пользовательским интерфейсом для Elastic Stack. Компоненты Logstash и Beats позволяют передавать данные в Elastic Stack. X-Pack предоставляет мощный функционал: можно настраивать мониторинг, добавлять различные уведомления, устанавливать параметры безопасности для подготовки вашей системы к эксплуатации. Поскольку Elasticsearch является ядром Elastic Stack

Elasticsearch — высокомасштабируемая распределенная поисковая система полнотекстового поиска и анализа данных, работающая в режиме реального времени. Утилита позволяет хранить, искать и анализировать большие объемы данных. Обычно используется в качестве базового механизма/технологии, помогая приложениям со сложными функциями поиска. Elasticsearch представляет собой основной компонент Elastic Stack.

Elasticsearch как сердце Elastic Stack играет основную роль в поиске и анализе данных. Она построена на уникальной технологии — Apache Lucene. Благодаря этому Elasticsearch в корне отличается от традиционных решений для реляционных баз данных или NoSQL. Ниже перечислены основные преимущества использования Elasticsearch в качестве хранилища данных:

- неструктурированность, документоориентированность;
- возможность поиска;
- возможность анализа данных;
- поддержка пользовательских библиотек и REST API;
- легкое управление и масштабирование;
- работа в псевдореальном времени;
- высокая скорость работы;
- устойчивость к ошибкам и сбоям.

Обзор компонентов Elastic Stack

Некоторые компоненты универсальны, их можно применять без Elastic Stack или других инструментов.

Elasticsearch

Elasticsearch хранит все ваши данные, предоставляет возможности поиска и анализа в масштабируемом виде. Мы уже рассматривали преимущества и причины использования Elasticsearch. Вы можете работать с Elasticsearch без каких-либо других компонентов, чтобы оснастить свое приложение инструментами для поиска и анализа данных.

Чтобы работать с реляционными базами данных, нужно разбираться в таких понятиях, как строки, столбцы, таблицы и схемы. Elasticsearch и другие хранилища, ориентированные на документы, работают по иному принципу. Система Elasticsearch имеет четкую ориентацию на документы. Лучше всего для нее подходят JSON-документы. Они организованы с помощью различных типов и индексов. Далее мы рассмотрим ключевые понятия Elasticsearch:

- индекс;
- тип;
- документ;
- кластер;
- узел;
- шарды и копии;
- разметку и типы данных;
- обратный индекс.

Индекс — это контейнер, который в Elasticsearch хранит документы одного типа и управляет ими. Индекс может содержать документы одного типа,

Индексы в Elasticsearch приблизительно аналогичны по структуре базе данных в реляционных базах данных. Продолжая аналогию, тип в Elasticsearch соответствует таблице, а документ — записи в ней.

Тип

Типы помогают логически группировать или организовывать однотипные документы по индексам.

Обычно документы с наиболее распространенным набором полей группируются под одним типом. Elasticsearch не требует наличия структуры, позволяя вам хранить любые документы JSON с любым набором полей под одним типом. На практике следует избегать смешивания разных сведений в

одном типе, таких как «клиенты» и «продукты». Имеет смысл хранить их в разных типах и с разными индексами.

Документ

Как уже было сказано, JSON-документы лучше всего подходят для использования в Elasticsearch. Документ состоит из нескольких полей и является базовой единицей информации, хранимой в Elasticsearch. Например, у вас может быть документ, соответствующий одному продукту, одному клиенту или одной позиции заказа.

Документы содержат несколько полей. В документах JSON каждое поле имеет определенный тип. В примере с каталогом продуктов, который мы видели ранее, были поля sku, title, description, price и др. Каждое поле и его значение можно увидеть как пару «ключ — значение» в документе, где ключ — это имя поля, а значение — значение поля.

Узел

Elasticsearch — распределенная система. Она состоит из множества процессов, запущенных на разных устройствах в сети и взаимодействующих с другими процессами. В главе 1 мы скачали, установили и запустили Elasticsearch. Таким образом мы запустили так называемый единичный узел кластера Elasticsearch.

Узел Elasticsearch — это единичный сервер системы, который может быть частью большого кластера узлов. Он участвует в индексировании, поиске и выполнении других операций, поддерживаемых Elasticsearch. Каждому узлу Elasticsearch в момент запуска присваиваются уникальный идентификатор и имя.

Каждому узлу Elasticsearch соответствует основной конфигурационный файл, который находится в подкаталоге настроек. Формат файла YML (полное название — YAML Ain't Markup Language). Вы можете использовать этот файл для изменения значений по умолчанию, таких как имя узла, порты, имя кластера.

На базовом уровне узел соответствует одному запущенному процессу Elastic-search. Он отвечает за управление соответствующей ему частью данных.

Кластер

Кластер содержит один или несколько индексов и отвечает за выполнение таких операций, как поиск, индексирование и агрегации. Кластер формируется одним или несколькими узлами. Любой узел Elasticsearch всегда является частью кластера, даже если это кластер единичного узла. По умолчанию каждый узел пытается присоединиться к кластеру с именем Elasticsearch. Если вы запускаете несколько узлов внутри одной сети без изменения параметра cluster.name в файле config/elasticsearch.yml, они автоматически объединяются в кластер.

Кластер состоит из нескольких узлов, каждый из которых отвечает за хранение своей части данных и управление ею. Один кластер может хранить один или несколько индексов. Индекс логически группирует разные типы документов.

Шарды и копии

Шарды помогают распределить индекс по кластеру. Они распределяют документы из одного индекса по различным узлам. Объем информации, который может храниться в одном узле, ограничивается дисковым пространством, оперативной памятью и вычислительными возможностями этого узла. Шарды помогают распределять данные одного индекса по всему кластеру и тем самым оптимизировать ресурсы кластера.

Процесс разделения данных по шардам называется шардированием. Это неотъемлемая часть Elasticsearch, необходимая для масштабируемой и параллельной работы с выполнением оптимизации:

- дискового пространства по разным узлам кластера;
- вычислительной мощности по разным узлам кластера.

Распределенные системы наподобие Elasticsearch приспособлены к работе даже при неполадках оборудования. Для этого предусмотрены реплики шардов, или копии. Каждый шард индекса может быть настроен таким образом, чтобы у него было некоторое количество копий или не было ни одной. Реплики шардов — это дополнительные копии оригинального или первичного шарда для обеспечения высокого уровня доступности данных.

Разметка и типы данных

Elasticsearch — неструктурированная система, благодаря чему в ней можно хранить документы с любым количеством полей и типов полей. В реальности данные никогда не бывают абсолютно бесструктурными. Всегда есть некий набор полей, общий для всех документов этого типа. Фактически типы внутри индексов должны создаваться на основе общих полей. Обычно один тип документов внутри индекса содержит несколько общих полей.

Типы данных

Elasticsearch поддерживает широкий набор типов данных для различных сценариев хранения текстовых данных, чисел, булевых, бинарных объектов, массивов, объектов, вложенных типов, геоточек, геоформ и многих других специализированных типов данных, например адресов IPv4 и IPv6. В документе каждое поле имеет ассоциированный тип данных.

Logstash

Утилита Logstash помогает централизовать данные, связанные с событиями, такие как сведения из файлов регистрации (логов), разные показатели (метрики) или любые другие данные в любом формате. Она может выполнить обработку данных до того, как сформировать нужную вам выборку. Это ключевой компонент Elastic Stack, который используется для сбора и обработки ваших контейнеров данных.

Logstash — компонент на стороне сервера. Его цель — выполнить сбор данных из обширного количества источников ввода в масштабируемом виде, обработать информацию и отправить ее по месту назначения. По умолчанию преобразованная информация поступает в Elasticsearch, но вы можете выбрать один из многих других вариантов вывода. Архитектура Logstash основана на плагинах и легко расширяется. Поддерживаются три вида плагинов: ввода, фильтрации и вывода.

Kibana

Kibana — инструмент визуализации для Elastic Stack, который поможет вам наглядно представить данные в Elasticsearch. Его также часто называют окном в Elastic Stack. В Kibana предлагается множество вариантов визуали-

заций, таких как гистограмма, карта, линейные графики, временные ряды и др. Вы можете создавать визуализации буквально парой щелчков кнопкой мыши и исследовать свои данные в интерактивном виде. Кроме того, есть возможность создавать красивые панели управления, состоящие из различных визуализаций, делиться ими, а также получать высококачественные отчеты.

В Kibana также предусмотрены инструменты для управления и разработки. Вы можете управлять настройками X-Pack для обеспечения безопасности в Elastic Stack, а с помощью инструментов разработчика создавать и тестировать запросы REST API.

Kibana Console представляет собой удобный редактор, который поддерживает функцию автозавершения и форматирования запросов во время их написания.

Что такое REST API? REST означает Representational State Transfer. Это архитектурный стиль для взаимодействия систем друг с другом. REST развивался вместе с протоколом HTTP, и почти все системы, основанные на REST, используют HTTP как свой протокол. HTTP поддерживает различные методы: GET, POST, PUT, DELETE, HEAD и др. Например, GET предназначен для получения или поиска чего-либо, POST используется для создания нового ресурса, PUT может применяться для создания или обновления существующего ресурса, а DELETE — для безвозвратного удаления.

Elastic Cloud

Elastic Cloud — это облачный сервис по управлению компонентами Elastic Stack, предоставляемый компанией Elastic (https://www.elastic.co/) — автором и разработчиком Elasticsearch и других компонентов Elastic Stack. Все компоненты продукта (помимо X-Pack и Elastic Cloud) созданы на базе открытого исходного кода. Компания Elastic обслуживает все компоненты Elastic Stack, проводит тренинги, выполняет разработку и предоставляет облачные сервисы.

Помимо Elastic Cloud, есть и другие облачные решения, доступные для Elastic-search, например Amazon Web Services (AWS). Основное преимущество Elastic Cloud в том, что он создан и обслуживается авторами Elasticsearch и других компонентов Elastic Stack.

Как вы можете видеть, Elasticsearch и Elastic Stack можно использовать для широкого спектра задач. Elastic Stack — это платформа с расширенным набором инструментов для создания комплексных решений поиска и аналитики. Она подходит для разработчиков, архитекторов, бизнес-аналитиков и системных администраторов. Вполне возможно создать решение на базе Elastic Stack, почти не прибегая к написанию кода, исключительно за счет изменения конфигурации. В то же время система Elasticsearch очень гибкая, следовательно, разработчики и программисты могут строить мощные приложения благодаря обширной поддержке языков программирования и REST API.

Sphinxsearch

Еще одна полнотекстовая поисковая система для больших данных – Sphinxsearch.

Sphinxsearch (от *SQL Phrase Index*) распространяется по лицензии GNU GPL либо, для версий 3.0+ без исходных кодов. Отличительной особенностью является высокая скорость индексации и поиска, а также интеграция с существующими СУБД (MySQL, PostgreSQL) и API для распространённых языков веб-программирования (официально поддерживаются PHP, Python, Java; существуют реализованные сообществом API для Perl, Ruby, .NET и C++).

Официальный сайт системы – http://sphinxsearch.com/.

Система Sphinxsearch обладает такими особенностями:

- Высокая скорость индексации (до 10-15 МБ/сек на каждое процессорное ядро);
- Высокая скорость поиска (до 150—250 запросов в секунду на каждое процессорное ядро с 1 000 000 документов);
- Большая масштабируемость (крупнейший известный кластер индексирует до 3 000 000 000 документов и поддерживает более 50 миллионов запросов в день);
- Поддержка распределенного поиска;
- Поддержка нескольких полей полнотекстового поиска в документе (до 32 по умолчанию);

- Поддержка нескольких дополнительных атрибутов для каждого документа (то есть группы, временные метки и т. д.);
- Поддержка однобайтовых кодировок и UTF-8;
- Поддержка морфологического поиска — имеются встроенные модули для английского, русского и чешского языков; доступны модули для французского, испанского, португальского, итальянского, румынского, немецкого, голландского, шведского, норвежского, датского, финского, венгерского языков;
- Нативная поддержка существующих СУБД PostgreSQL и MySQL, поддержка ODBC совместимых баз данных (MS SQL, Oracle и т. д.).

В 2017 году команда Manticore Software сделала форк Sphinxsearch 2.3.2, который назвали Manticore Search. По словам разработчиков, менеджмент системы Sphinxsearch не справлялся с сопровождением системы, а именно, не исправлялись обнаруженные ошибки, не реализовывались объявленные возможности, диалог пользователей с разработчиками Sphinxsearch был затруднен. Sphinx версии 3 уже можно воспринимать как проприетарное решение для ограниченного круга пользователей. Фактически новая версия системы (Manticore Search) решила монгие из этих проблем, в том числе, обеспечена поддержка кода в целом, организовано современнное взаимодействие с пользователями Sphinxsearch и Manticore, реализаованы такие возможности, присущие, в частности, Elasticsearch: репликация, auto id, JSON інтерфейс, возможность создать/удалить индекс на лету, наличие хранилища документов, развитых real-time индексов.

Neo4j

Neo4j — графовая система управления базами данных с открытым исходным кодом, языке Java, с поддержкой транзакции (ACID). По состоянию на 2015 год считается самой распространённой графовой СУБД [Робинсон, 2016]. Разработчик — американская компания Neo Technology, разработка ведётся с 2003 года.

Данные хранит в собственном формате, специализированно приспособленном для представления графовой инфор-

мации, такой подход в сравнении с моделированием графовой базы данных средствами реляционной СУБД позволяет применять дополнительную оптимизацию в случае данных с более сложной структурой. Также утверждается о наличии специальных оптимизаций для SSD-накопителей, при этом для обработки графа не требуется его помещение целиком в оперативную память вычислительного узла, таким образом, возможна обработка достаточно больших графов.

Основные области применения: социальные сети, системы предоставления рекомендаций, выявление мошенничества, картографические системы.

Терминология графовых баз данных

- graph database, графовая база данных — база данных построенная на графах — узлах и связях между ними
- Cypher — язык для написания запросов к базе данных Neo4j (примерно, как SQL в MYSQL)
- node, нода — объект в базе данных, узел графа. Количество узлов ограниченно 2 в степени 35 ~ 34 биллиона
- node label, метка ноды — используется как условный «тип ноды». Например, ноды типа movie могут быть связанны с нодами типа actor. Метки нод — регистрозависимые, причем *Cypher не выдает ошибок, если набрать не в том регистре название.
- relation, связь — связь между двумя нодами, ребро графа. Количество связей ограниченно 2 в степени 35 ~ 34 биллиона
- relation identirfier, тип связи — в Neo4j у связей. Максимальное количество типов связей 32767
- properties, свойства ноды — набор данных, которые можно назначить ноде. Например, если нода — это товар, то в свойствах ноды можно хранить id товара из базы MySQL
- node ID, ID нода — уникальный идентификатор ноды. По умолчанию, при просмотрах результата отображается именно этот ID.

Сохранение данных в Neo4j

Файл nodestore.db содержит определенный размер записей, содержащих информацию о ноде:

1. Метка, которая показывает, запись активна;
2. Указатель на первое отношение, которое содержит данная нода;
3. Указатель на первую свойство, которое содержит данная нода.

Нода не содержит собственного идентификатора. Так как каждая запись в nodestore.db занимает одинаковое количество места, можно рассчитать указатель на ноду.

Файл relationshipstore.db также содержит записи одинакового размера, которые описывают отношения, но они состоят из следующих элементов:

1. Метка, которая показывает, запись активна;
2. Указатель на ноду, которая содержит это отношение;
3. Указатель на ноду, к которой это отношение направлено;
4. Вид отношения;
5. Указатель на отношение, которое стоит впереди (в пределах данной ноды);
6. Указатель на отношение, которое стоит сзади (в пределах данной ноды);
7. Указатель на отношение, которое стоит впереди (в пределах Ноды, в которой это отношение направлено);
8. Указатель на отношение, которое стоит сзади (в пределах Ноды, в которой это отношение направлено);
9. Указатель на первое свойство данного отношения.

Как модель данных выбран ориентированый граф свойств:

- Содержит узлы (nodes) и связи (relationships).
- Узлы имеют свойства (properties). Узлы можно рассматривать как документы, содержащие свойства в виде пар ключ-значение.
- Узлы могут быть обозначены одной или несколькими метками (labels). Метки группируют узлы, указывая роль, которую они играют в наборе данных.

Одному узлу можно приписывать несколько меток (поскольку узлы могут играть несколько разных ролей в разных доменах). Связи связывают узлы и структурируют граф. Связи именуемые (всегда имеют одно имя) и направлены (всегда имеют направление, начальный и конечный узлы).Связи также могут содержать свойства. Это позволяет ввести дополнительные метаданные в графу алгоритмы, добавить дополнительную семантику связям, ограничивать запросы в режиме реального времени.

Основные транзакционные возможности — поддержка ACID и соответствие спецификациям JTA, JTS и XA. Интерфейс программирования приложений для СУБД реализован для многих языков программирования, включая Java, Python, Clojure, Ruby, PHP, также реализовано API в стиле REST. Расширить программный интерфейс можно как с помощью серверных плагинов, так и с помощью неуправляемых расширений (*unmanaged extensions*); плагины могут добавлять новые ресурсы к REST-интерфейсу для конечных пользователей, а расширения позволяют получить полный контроль над программным интерфейсом, и могут содержать произвольный код, поэтому их следует использовать с осторожностью.

В СУБД используется Cypher — декларативная язык запросов к графам. Синтаксис этого языка похож на синтаксис SQL. Поддерживаются операции по созданию, выборки, обновления, удаления данных. Cypher описывает графы, используя спецификацию по образцу — используется простая форма ASCII-графики, пользователь рисует часть графа, его интересует, с помощью ASCII символов; вершины берутся в скобки, их метки прописываются после «:»; для создания нескольких узлов их следует перечислить через »,«; связи отражаются стрелками (-> и <-), а названия связей указываются внутри квадратных скобок после «:»; свойства узлов и связей (пары ключ-значение) прописываются в фигурных скобках.

Язык запросов Cypher — самый распространенный язык запросов к графовым базам данных, что обусловлено его использованием в СУБД Neo4j . Cypher является декларативным языком и позволяет создавать, обновлять и удалять вершины, ребра, метки и свойства, а также управлять индексами и ограничениями. Для извлечения данных из хранилища испо-

льзуется запрос, содержащий шаблон фильтрации, позволяющий получать:

- (n)→(m) — все направленные ребра из вершины n в вершину m;
- (n:Person) — все вершины с меткой Person;
- (n:Person:Russian) — все вершины, имеющие обе метки Person и Russian;
- (n:Person {name:{value}}) — все вершины с меткой Person и отфильтрованные по дополнительному свойству;
- (n:Person) → (m) — ребра между вершинами n с меткой Person и m;
- (n)--(m) — все ненаправленные ребра между вершинами n и m.

Запросы в Neo4j можно делать и другими способами, например, напрямую через Java API и на языке Gremlin[en], созданном в проекте с открытым исходным кодом TinkerPop. Cypher является не только языком запросов, но и языком манипулирования данными, так как предоставляет функции CRUD для графового хранилища.

Gephi

Gephi - это пакет программного обеспечения с открытым кодом для анализа и визуализации графов (сетей).

Gephi (https://gephi.org/) - это в настоящее время самая популярная программа визуализации и анализа сетей и графов («сетевых графов»). Gephi обеспечивает быструю компоновку, эффективную фильтрацию и интерактивное исследование данных, а также является одним из лучших вариантов для визуализации крупномасштабных сетей. Gephi - это мультиплатформенное программное обеспечение, которое распространяется с открытым кодом согласно лицензиям CDDL 1.0 и GNU General Public License v3. По адресу https://gephi.org/ доступны версии для Mac OS X, Windows и Linux исходных кодов.

Gephi активно используется в целом ряде академических исследовательских проектов, в частности социологических; также быстро получил популярность среди журналистов. Сейчас его пользовательское среду значительно расширился - с помощью этого пакета можно заниматься любой темой сете-

вого анализа. Gephi использовался, среди прочего, для визуализации глобальной связности контента New York Times и изучения сетевого трафика Twitter во время социальных волнений; Gephi вдохновлял создания LinkedIn InMaps и был использован для визуализации целой сети Truthy.

Gephi позволяет обрабатывать графу структуры достаточно больших объемов (до 1 млн. Узлов) на персональном компьютере за счет ефективних алгоритмов.

Разработчики Gephi описывают эту программу как "как Photoshop, но для данных".

Программа включает в себя мнгожину различных алгоритмов компоновки (заключение графиков на плоскости) и позволяет настраивать цвета, размеры и метки в графах. Gephi является интерактивным программным обеспечением и предоставляет средства для выявления сообществ, а также предоставляется возможность расчета кратчайших путей или относительного расстояния от любого узла к данному узлу. Плагины от Gephi позволяют расширять ее функциональность и добавлять новые алгоритмы, макеты и инструменты измерений. Gephi имеет многопоточную схему обработки данных и таким образом, позволяет выполнять несколько видов анализа одновременно.

Интерфейс пользователя системы Gephi включает три основных раздела (окна):

- «Лаборатория данных»: здесь хранятся все исходные данные о сети, а также дополнительные расчетные значения;
- «Обработка данных»: здесь происходит большая часть операций пользователя, в частности, ручное редактирование сетей, тестирование макетов, установки фильтров;
- «Предварительный просмотр»: здесь уточняется форма вывода графу, как правило, с помощью набора инструментов граф дорабатывается, в том числе, и с эстетической точки зрения. В этом же окне реализован вызов экспорта графа в форматы PDF, PNG и SVG.

Программа включает в себя мнгожину различных алгоритмов компоновки (заключение графиков на плоскости) и позволяет настраивать цвета, размеры и метки в графах.

Gephi является интерактивным программным обеспечением и предоставляет средства для выявления сообществ, а также предоставляется возможность расчета кратчайших путей или относительного расстояния от любого узла к данному узлу. Плагины от Gephi позволяют расширять ее функциональность и добавлять новые алгоритмы, макеты и инструменты измерений. Gephi имеет многопоточную схему обработки данных и таким образом, позволяет выполнять несколько видов анализа одновременно.

Эти три основных раздела охватывают множество вкладок, которые позволяют пользователю реализовывать отдельные функции. Ниже рассматривается каждое из основных и вторичных окон - разделов и вкладок.

При анализе больших и плотных сетей, быстрое компоновки (упорядочение узлов графов) является узким местом, поскольку большинство сложных алгоритмов компоновки является требовательными к параметрам процессора, памяти и времени выполнения. В то же время, Gephi поставляется с эффективными алгоритмами компоновки, такими как Yifan-Hu, Force-directed. В частности, алгоритм Yifan-Hu является идеальным вариантом для применения после других, более быстрых и грубых алгоритмов. В то время, как большинство из предложенных в Gephi методов могут выполняться в течение допустимого времени, сочетание, например, OpenOrd и Yifan-Hu, дает наиболее качественные визуальные представления. Конечно, правильная параметризация любого алгоритма компоновки может влиять как на работе, так и на результат визуализации.

Gephi позволяет загружать данные сетей в форматах GEXF, GDF, GML, GraphML, Pajek (NET), GraphViz (DOT), CSV, UCINET (DL), Tulip (TPL), Netdraw (VNA) и таблиц Excel. Кроме того, Gephi позволяет экспортировать данные сетей в форматах JSON, CSV, Pajek (NET), GUESS (GDF), Gephi (GEFX), GML и GraphML. Благодаря этому Gephi может взаимодействовать с другими системами анализа и визуализации графов.

2.6. Математические основы

В данной главе основное внимание уделено распознаванию информационных операций на основе изучения динамиче-

ских свойств информационных потоков в глобальных компьютерных сетях, в частности, в сети Интернет.

Для исследования информационных потоков в Интернете, т.е. потока сообщений, которые публикуются на страницах веб-сайтов, в социальных сетях, блогах, и т.п., должен применяться современный инструментарий. Так известные методы обобщения информационных массивов (классификация, фазовое укрупнение, кластерный анализ и т.д.) уже не всегда пригодны даже для адекватного количественного отражения процессов, происходящих в информационном пространстве [Lande, 2007].

Количественный анализ динамики информационных потоков, которые генерируются в Интернете, становится сегодня одним из наиболее информативных методов исследования актуальности тех или иных тематических направлений. Эта динамика обусловлена разнообразными качественными факторами, многие из которых не поддаются точному описанию. Однако общий характер временной зависимости количества тематических публикаций в сети Интернет все же допускает построение математических моделей, их исследование, прогнозирование. Наблюдения временных зависимостей объемов сетевых информационных потоков убедительно свидетельствуют о том, что механизмы их генерации и распространения, очевидно, связаны со сложными нелинейными процессами. Именно этой теме посвящена данная глава.

Для анализа временных рядов, которые отображают зависимость объемов информационных потоков от времени, используют разнообразные методы и подходы. При этом оказывается, что все эти подходы взаимосвязаны и более того, ключевую роль играет понятие корреляции. Изложение построено вокруг схемы показанной на Рис. 13, причем особое внимание уделено взаимосвязям.

Рис. 13 – Взаимосвязи между подходами к анализу временных рядов

2.6.1. Временные ряды

Временной ряд – это набор наблюдаемых значений упорядоченных по времени. Далее будут рассматриваться дискретные временные ряды, значения которых фиксировались через равные промежутки времени. Будем обозначать такой временной ряд $x_1, x_2, \dots, x_T$ или коротко $\{x_t\}_{t=1}^{T}$ подразумевая, что фиксирование значений ряда происходило через равный промежуток времени h: $t_0, t_0 + h, t_0 + 2h, \dots, t_0 + (T-1)h$.

Если значения временного ряда однозначно задаются некоторым математическим соотношением (таким как, например, $x_t = A \cdot sin(vt)$), то такой ряд является детерминированным. Если значения временного ряда можно описать только в терминах вероятностного распределения, то речь идет о статистическом временном ряде. Такие ряды и будут рассматриваться далее. Анализируя временные ряды, мы будем рассматривать их как реализацию стохастического процесса.

В качестве примеров далее будут использоваться три временных ряда, которые были получены с помощью популярного сетевого сервиса GoogleTrends. Эти временные ряды отображают уровень интереса к Дональду Трампу, Хилари Клинтон и «русским хакерам» с августа 2016 года по апрель 2017 года. Временные ряды, получаемые с помощью GoogleTrends, показывают динамику популярности поискового запроса.

Максимальная точка на графике равна 100 и соответствует дате, когда запрос был наиболее популярен, а остальные точки на графике определяются в процентном соотношении к максимуму. Все три временных ряда показаны на Рис. 14. Для простоты ссылок на данные ряды в дальнейшем обозначим их T (Д. Трамп), K (Х. Клинтон), X («русские хакеры»).

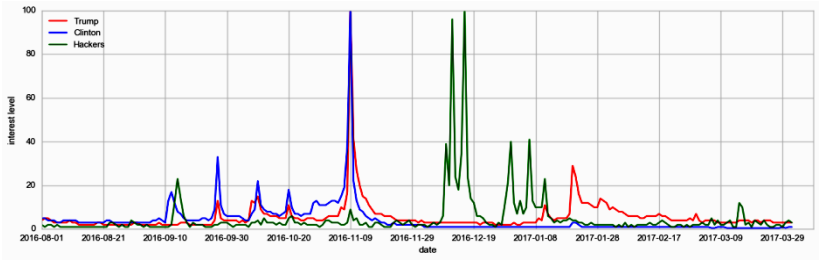


Рис. 14 – Временные ряды, которые отображают интерес к Дональду Трампу (T), Хилари Клинтон (K) и «русским хакерам» (X) с 1 августа 2016 года по 1 апреля 2017 из GoogleTrends.

В некоторых случаях полезно рассмотреть более гладкую версию исходного временного ряда. Сглаживание помогает выявить существенные тенденции в динамике ряда, скрыв при этом шум и различные особенности, которые проявляются при небольших масштабах. Существуют разнообразные методы сглаживания. Наиболее простой способ сглаживания – это вычисление скользящего среднего. Простое скользящее среднее равно среднему арифметическому значению элементов ряда из интервала заданной длины, а именно

$$SMA_t = \frac{1}{w}\sum_{i=0}^{w-1} x_{t-i},$$

где w – ширина сглаживающего интервала (количество элементов, по которым рассчитывается среднее), SMA_t – значение простого скользящего среднего в точке t. Полученное значение SMA_t относится к середине сглаживающего интервала, поэтому сглаженный ряд y_t может быть определен как $y_t = SMA_{t+\left[\frac{w}{2}\right]}$.

При использовании сглаживания скользящим средним, чем больше ширина сглаживающего интервала, тем более

гладкой получится функция. На Рис. 15 показано как выглядит сглаженный ряд T при увеличении значения w.

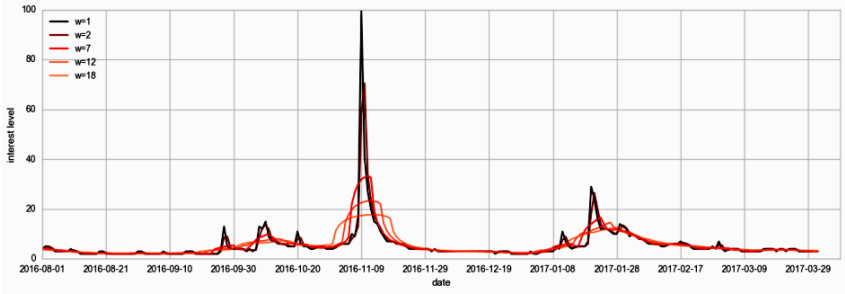


Рис. 15 – Исходный ряд T и сглаженный простым скользящим средним с шириной сглаживающего интервала 2, 7, 12, 18.

Результаты сглаживания ряда можно продемонстрировать на графике, у которого ось абсцисс соответствует временной оси, а вдоль оси ординат отложена ширина сглаживающего интервала. На графике показаны значения $y_t^{(w)}$ – то есть элементы сглаженного ряда в точке t при использовании интервала ширины w (Рис. 16).

При вычислении простого скользящего среднего все точки, которые попали в сглаживающий интервал, имеют одинаковый вес. Естественно, что можно использовать не равные веса. Таким образом, приходим к определению взвешенного скользящего среднего

$$WMA_t = \frac{1}{w}\sum_{i=0}^{w-1} a_i x_{t-i},$$

где $\sum_{i=0}^{w-1} a_i = 1$.

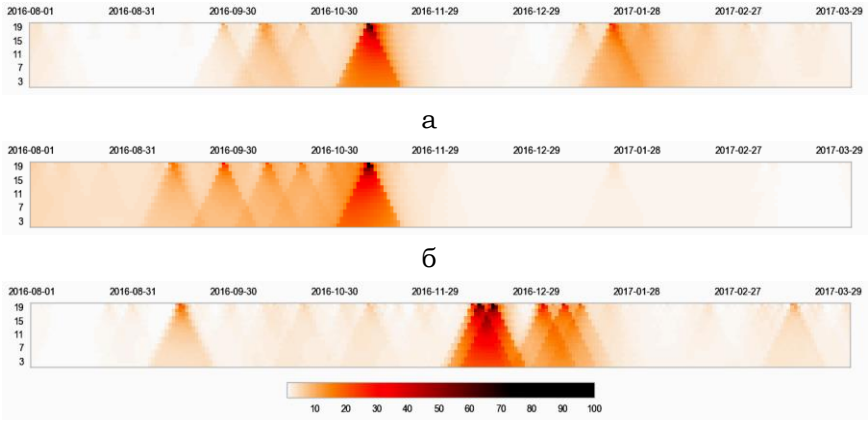


Рис. 16 – Значения сглаженных простым скользящим средним временных рядов T (а), K (б) и X (в) в зависимости от ширины сглаживающего интервала. Вдоль оси абсцисс отложено время, а вдоль оси ординат – ширина интервала.

Другой часто используемый метод сглаживания рядов – это ***экспоненциальное сглаживание***. Предыдущие значения ряда учитываются с экспоненциально убывающими весами. Будем обозначать элементы сглаженного ряда y_t, и сразу определим $y_0 = x_0$. Следующие элементы ряда y_t получают по рекурсивной формуле

$$y_t = \alpha x_t + (1 - \alpha) y_{t-1},$$

где $0 < \alpha < 1$ – коэффициент сглаживания. Очевидно, что при $\alpha = 1$ получаемый ряд y_t совпадает с исходным x_t. Таким образом, если значение α близко к 1, то наибольший вес при определении y_t присваивается соответствующему x_t, а предыстория ряда «мало значит». С другой стороны, если бы α равнялось 0, то весь ряд y_t сгладился бы до одного значения $y_t = y_0$. То есть при α близком к 0 предыстория ряда учитывается с большим весом, чем текущее значение.

На Рис. 17 показан ряд T, а также соответствующие сглаженные ряды при различных значениях параметра α.

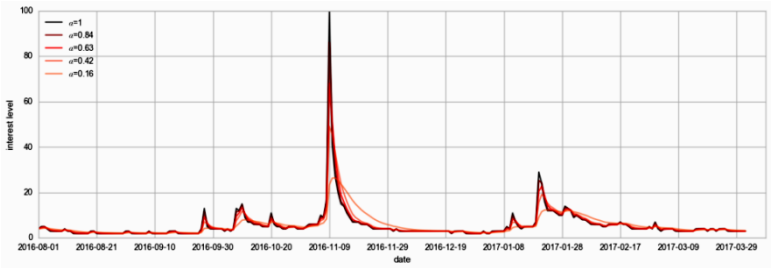


Рис. 17 – Исходный ряд T и сглаженный с помощью экспоненциального сглаживания с параметром, равным 0.84, 0.63, 0.42, 0.16

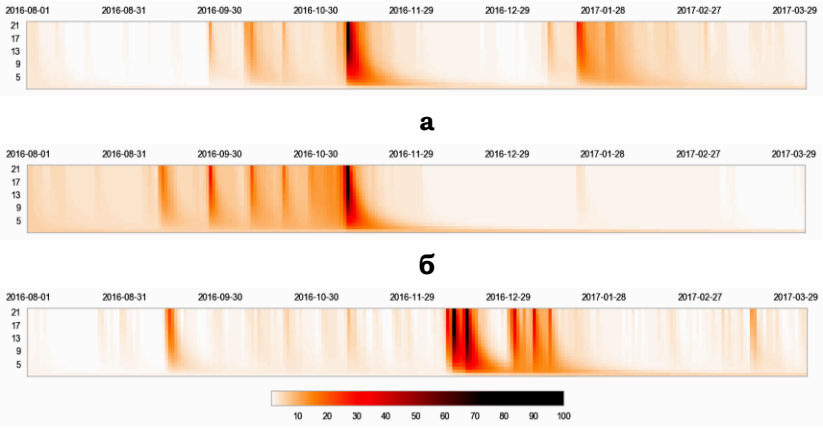
Как и в случае с простым скользящим средним продемонстрируем результаты сглаживания ряда на графике. В данном случае вдоль оси ординат отложим параметр α (Рис. 18). На графике показаны значения $y_t^{(\alpha)}$ – значение в точке t сглаженного с параметром α исходного ряда.

В качестве примеров мы рассматриваем временные ряды T, K, и X, у которых есть недельная периодичность. Это характерное свойство многих процессов в информационном пространстве. Известно, что публикация новостных сообщений часто происходит с недельной периодичностью, а также активность пользователей разнится в будни и выходные дни.

Для того чтобы исключить периодическую компоненту из рядов, сгладим их с помощью простого скользящего среднего с интервалом шириной 7 (число дней в неделе) в соответствии с формулой:

$$x_t^{New} = \frac{x_{t-3} + x_{t-2} + x_{t-1} + x_t + x_{t+1} + x_{t+2} + x_{t+3}}{7},$$

где x_t – исходные значения ряда, x_t^{New} – новое значение ряда в момент времени t. На Рис. 19 показаны сглаженные временные ряды.



а

б

в

Рис. 18 – Значения экспоненциального сглаженных временных рядов T (а), K (б) и X (в) в зависимости от параметра α

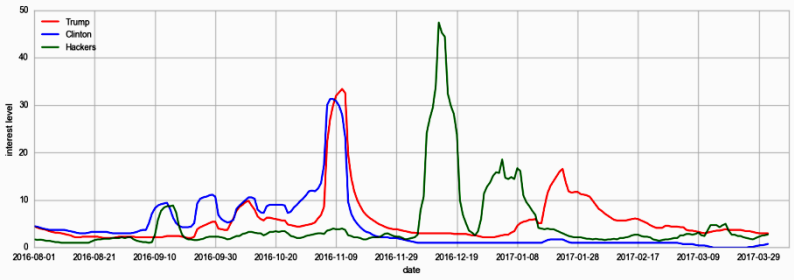


Рис. 19 – Временные ряды T, K и X, сглаженные с помощью простого скользящего среднего с интервалом длины 7

2.6.2. Корреляционный анализ

Многие методы исследования временных рядов базируются на некотором предположении о статистическом равновесии или постоянстве. Одним из таких полезных предположений является стационарность [Box 2015].

Временной ряд называется строго стационарным или стационарным в узком смысле, если его статистические свойства не изменяются со временем. Формально, если совместное распределение случайных величин $x_t, x_{t+1}, \dots, x_{t+n}$ совпадает с распределением $x_{t+k}, x_{t+k+1}, \dots, x_{t+k+n}$ при любых целых значениях сдвига k, то временной ряд $\{x_t\}_{t=1}^{T}$ называется строго стационарным. У стационарных временных рядов постоянное математическое ожидание

$$\mu = Ex_t$$

и дисперсия

$$\sigma^2 = Var(x_t) = E(x_t - Ex_t)^2.$$

При этом значения μ и σ^2 можно оценить как выборочное среднее

$$\hat{\mu} = \bar{x} = \frac{1}{T}\sum_{t=1}^{T} x_t$$

и выборочную дисперсию

$$\hat{\sigma}^2 = \frac{1}{T}\sum_{t=1}^{T}(x_t - \bar{x})^2.$$

Свойство стационарности также имеет большое значение при сравнении временных рядов. Линейная зависимость между двумя случайными величинами измеряется ковариацией. Для временных рядов определяют кроссковариационную функцию. По определению, кросс-ковариация с временной задержкой k между случайными процессами $\{x_t\}_{t=1}^{T}$ и $\{y_t\}_{t=1}^{T}$ равна

$$\gamma_{xy}(k,t) = Cov(x_t, y_{t+k}) = E\big[(x_t - \mu_x)(y_{t+k} - \mu_y)\big].$$

Из предположения о стационарности в узком смысле следует, что распределение пар величин x_t, y_{t+k} одинаково для произвольно значения t. Следовательно, ковариация между

величинами x_t и y_{t+k} не зависит от t, а зависит только от значения k, то есть $\gamma_{xy}(k,t) = \gamma_{xy}(k), \forall t$. Набор значений $\{\gamma_{xy}(k)\}$ образует кроссковариационную функцию.

Нормировав кроссковариационный коэффициент, получим кросскорреляционный коэффициент

$$\rho_{xy}(k) = \frac{Cov(x_t, y_{t+k})}{\sigma_x \sigma_y} = \frac{\gamma_{xy}(k)}{\sigma_x \sigma_y}.$$

Кросскорреляционная функция является мерой подобия между двумя временными рядами.

Чаще всего кроссковариационные и кросскорреляционные коэффициенты оценивают по формулам

$$\hat{\gamma}_{xy}(k) = \frac{1}{T}\sum_{t=1}^{T-k}(x_t - \bar{x})(y_{t+k} - \bar{y}), \qquad \hat{\rho}_{xy}(k) = \frac{\hat{\gamma}_{xy}(k)}{\hat{\gamma}_{xy}(0)}.$$

Заметим, что такие оценки справедливы для стационарных в узком смысле рядов, так как соответствующие коэффициенты не зависят от времени, а в общем случае это может и не выполнятся. Часто используют более слабое требование, чем стационарность в узком смысле, – стационарность в широком смысле.

Временной ряд $\{x_t\}_{t=1}^{T}$ стационарен в широком смысле, если его математическое ожидание не изменяется со временем, то есть $\forall t \; \exists Ex_t = \text{const}$ и ковариационная функция зависит только от разности аргументов $\text{Cov}(x_t, x_s) = K(t - s)$.

Так как в определении указано, что математическое ожидание постоянно и, легко заметить, что дисперсия также не изменяется со временем $Var(x_t) = Cov(x_t, x_t) = K(0) = const$, то в этом случае, как и для строго стационарных рядов, справедливы оценки (1) и (2).

На Рис. 20 показана иллюстрация к вычислению корреляции.

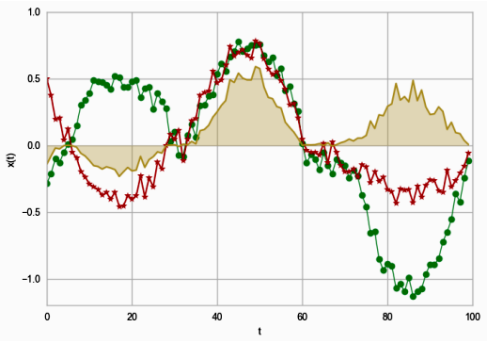


Рис. 20 – Иллюстрация к определению корреляции. Показаны два временных ряда

Рассматриваются два центрированных временных ряда. Для вычисления коэффициента корреляции нужно умножить соответствующие элементы рядов и вычислить их среднее значение. Результат умножения на Рис. 3.8 показан линией. Площадь затемненной области под линией с учетом знака равна коэффициенту ковариации между двумя рядами.

Для примера приведем оценку кросскорреляционных функций для рядов T, K, X. На Рис. 21 а показана корреляционная функция для рядов T и K. Вдоль оси абсцисс отложена временная задержка (лаг), вдоль оси ординат – оценка корреляционного коэффициента. Максимальное значение (приблизительно равное 0,8) функция достигает при временной задержке 0. То есть два временных ряда, связанные с интересом к Дональду Трампу и Хилари Клинтон, сильно коррелированны. На Рис. 21 б показана корреляционная функция для рядов T и X. Максимальное значение (приблизительно равное 0.7) функция достигает при временной задержке 34 дня. Это отвечает тому факту, что начиная с 13 декабря 2016 года (34 дня после выборов в США 8 ноября) резко возросло количество новостных сообщений про «русских хакеров».

Автокорреляция

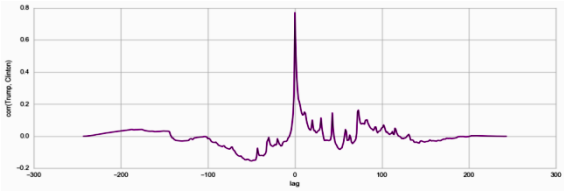
Можно подсчитать ковариацию не для двух различных рядов, а для одного ряда. Такая ковариация называется автоковариацией с временной задержкой или лагом k

$$\gamma_k = \mathrm{Cov}(x_t, x_{t+k}) = \mathrm{E}[(x_t - \mu)(x_{t+k} - \mu)].$$

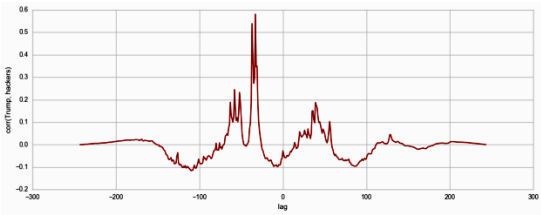
Набор величин γ_k, k = 0,1,2, … называется автоковариационной функцией, а их нормированное значение ρ_k, k = 0,1,2, … – автокорреляционной функцией

$$\rho_k = \frac{E[(x_t - \mu)(x_{t+k} - \mu)]}{\sqrt{E(x_t - \mu)^2 E(x_{t+k} - \mu)^2}} = \frac{Cov(x_t, x_{t+k})}{Var(x_t)} = \frac{\gamma_k}{\gamma_0}.$$

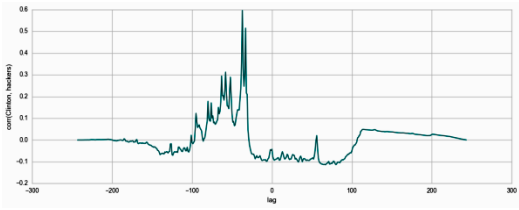
Автокорреляционная функция описывает зависимость между значениями случайного процесса в различные моменты времени (Рис. 22). На рисунке показан временной ряд и тот же самый ряд, сдвинутый на 10 значений вправо Затемненная область показывает вклад в значение автокорреляционного коэффициента с лагом 10.



а



б



в

Рис. 21 – Корреляционные функции для пар рядов Т и К (а), Т и Х (б), К и Х (в)

На Рис. 23 приведены автокорреляционные функции для рядов T (а), K (б), X (в). Вдоль оси абсцисс отложена временная задержка (лаг), вдоль оси ординат – автокорреляционный коэффициент. Затемненная область показывает стандартное отклонение для оценки автокорреляционного коэффициента.

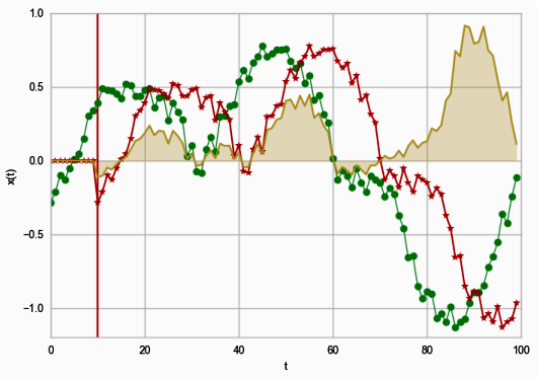


Рис. 22 – Иллюстрация к определению автокорреляции

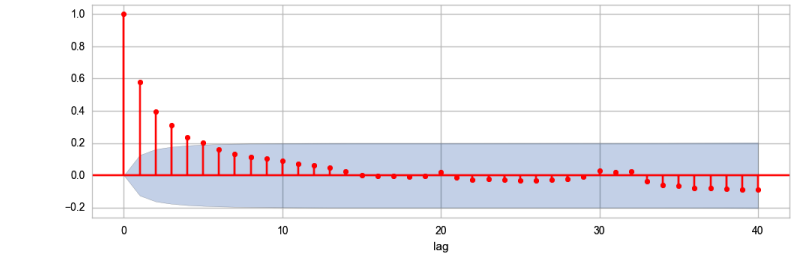
Чаще всего автоковариационные и автокорреляционные коэффициенты оценивают по формулам

$$\hat{\gamma}_k = \frac{1}{T}\sum_{t=1}^{T-k}(x_t - \bar{x})(x_{t+k} - \bar{x}), \qquad \hat{\rho}_k = \frac{\hat{\gamma}_k}{\hat{\gamma}_0}.$$

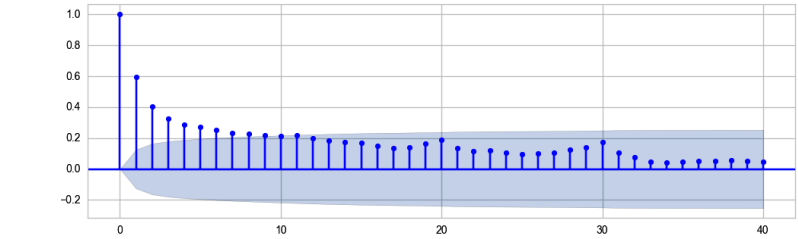
После вычисления оценок для автокорреляционных коэффициентов возникает вопрос: равны ли коэффициенты ρ_k нулю начиная с некоторого значения k? Для ответа на этот вопрос нужно сравнить значение оценки $\hat{\rho}_k$ с его стандартным отклонением. Если мы принимаем предположение, что $\rho_k = 0$, то стандартное отклонение оценки $\hat{\rho}_k$

$$se(\hat{\rho}_k) \cong \frac{1}{\sqrt{T}}.$$

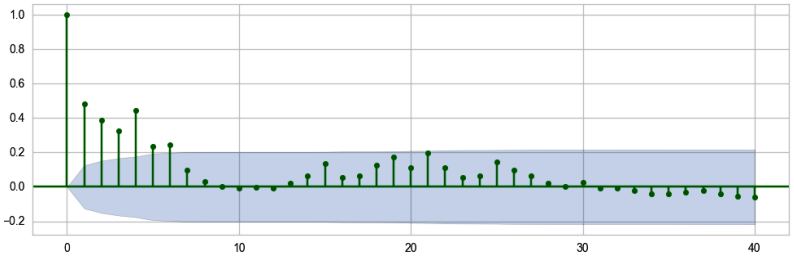
На практике часто используют эмпирическое правило, согласно которому автокорреляционные коэффициенты оценивают для временной задержки не более чем $T/4$.



а



б



в

Рис. 23 – Автокорреляционные функции для рядов T (а), K (б), X (в)

Определение автокорреляционной функции вводилось для стационарных временных рядов, но оценить ее значение можно для произвольного временного ряда. Для нестационарных временных рядов такая автокорреляционная функция убывает очень медленно.

2.6.3. Анализ Фурье

Классический анализ Фурье предоставляет возможность исследовать функцию во временной и частотной области. Суть перехода в частотную область состоит в том, что функ-

ция раскладывается на составляющие, которые являются гармоническими колебаниями с разными частотами. При этом каждой частоте соответствует коэффициент, который отображает амплитуду колебания на данной частоте. Если представить функцию графически во временной области, то получим информацию о том, как функция изменяется со временем. Если изобразить функцию в частотной области, то получим информацию о частотах, колебания на которых она содержит. Для этого используют прямое и обратное преобразование Фурье

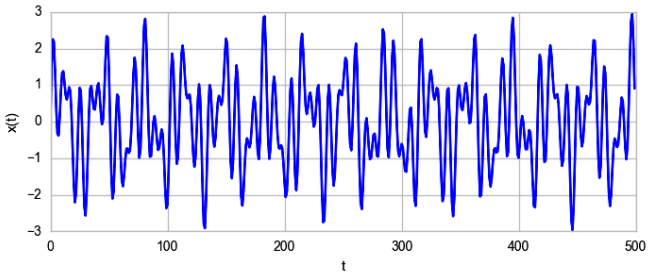
$$\hat{x}(v) = \int_{-\infty}^{\infty} x(t)e^{-i2\pi vt}dt,$$

$$x(t) = \int_{-\infty}^{\infty} \hat{x}(v)e^{i2\pi vt}dv.$$

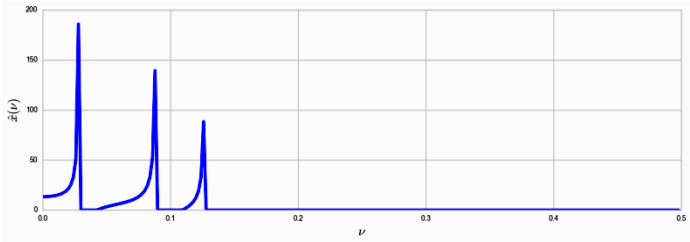
На Рис. 24 а показан пример функции, которая на самом деле является суммой трёх синусоид с разными периодами. Глядя только на график функции во временной области достаточно трудно понять, что она состоит из трёх гармонических колебаний и определить их периоды. На Рисунке 24 б показано преобразование Фурье для этой функции. Из графика в частотной области наглядно видно, что функция содержит колебания на трёх разных частотах.

Сегодня преобразование и спектры Фурье находят разнообразные применения в системах машинного обучения. Часто спектры Фурье используются в качестве обучающих параметров. Например, в [Rodrigues 2014] предложена модель прогнозирования временного ряда, в которой спектр Фурье вместе с некоторыми другими параметрами подается на вход нейронной сети.

Преобразования и спектры Фурье часто используются при распознавании речи. В [Alam 2014] специальные признаки, сформированные на основе преобразования Фурье, используются в системе распознавания речи с различными условиями обучения.



а



б

Рис. 24 – Функция, которая является суммой трёх синусоид с различными периодами (а) и оцененный спектр Фурье для этой функции (б)

Спектры Фурье также используются в качестве обучающих параметров для нейронных сетей в системах автоматического детектирования определенных событий в речи или на фоне шума [Sazonov 2010, Wang 2014]. Другой задачей в области распознавания является определение эмоциональной окраски речи. В [Wang 2015] предложена модель распознавания, в основе которой лежат определенные параметры Фурье, и демонстрируется эффективность использования таких параметров для идентификации различных эмоциональных состояний в голосовых сигналах.

В алгоритмах машинного обучения основанных на применения ядра, таких как машина опорных векторов (support vector machine), для аппроксимации ядер высокой размерности часто используют случайные признаки Фурье (random Fourier features). Такой подход был предложен в [Rahimi 2008] и основан на теореме Бохнера из гармонического анализа,

которая гарантирует, что при некоторых свойствах ядра его преобразование Фурье будет вероятностным распределением.

Преобразование Фурье можно воспринимать как определение корреляции между исходным сигналом и гармоническими функциями с различными частотами колебания. На Рис. 25 показана иллюстрация аналогичная с Рис. 20 и Рис. 22. Затемненная область показывает вклад в значение преобразования Фурье или амплитуды, которая соответствует данной частоте колебания.

Несмотря на свои преимущества и многочисленные приложения, преобразование Фурье является плохим методом для исследования функций, которые эволюционируют со временем. Для таких функций нужен некоторый способ оценивания спектра не по всей длине временного ряда, а по его различным частям. Примером такого подхода является оконное преобразование Гэбора

$$G(\nu, \tau, s) = \int_{-\infty}^{\infty} x(t) e^{-\frac{(t-\tau)^2}{s^2}} e^{-i2\pi\nu t} dt.$$

Временное окно $e^{-\frac{(t-\tau)^2}{s^2}}$ выделяет отрезок временного ряда с центром в точке τ и имеет ширину, которая определяется параметром s, что позволяет выделить часть исследуемого ряда.

При использовании преобразования Гэбора возникает проблема выбора ширины окна. Сделать оконную функцию зависящей от частоты так, чтобы для низких частот окно становилось шире, а высоких – уже, позволяет следующий класс преобразований, а именно вейвлет преобразование. Основное преимущество вейвлет преобразования состоит в том, что выделенный из временного ряда кусок анализируется с той степенью детальности, которая соответствует его масштабу.

2.6.4. Вейвлет-анализ

Вейвлет преобразование имеет корреляционную природу. В данном случае рассматривается корреляция исходной функции с функцией вейвлетом на разных масштабах. Для того чтобы такую процедуру всегда можно было выполнить и корреляционные коэффициенты были информативными,

вейвлет должен обладать определенными математическими свойствами.

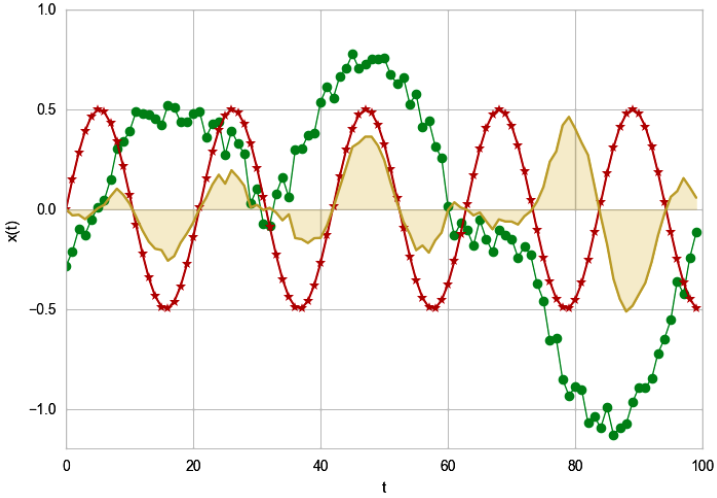


Рис. 25 – Иллюстрация к определению преобразования Фурье как вычисление корреляции между исходным сигналом и гармоническим колебанием.

Буквально слово вейвлет переводится как «маленькая волна» или «всплеск», и, как следует из названия, вейвлет хорошо локализирован во времени. С математической точки зрения, вейвлет – это функция $\psi(t)$, которая удовлетворяет следующим свойствам:

1. Функция $\psi(t)$ квадратично интегрируема ($\psi \in L^2(\mathbb{R})$) или, другими словами, имеет конечную энергию

$$E = \int_{-\infty}^{\infty} |\psi(t)|^2 dt < \infty.$$

2. Обозначим $\hat{\psi}(\lambda)$ преобразование Фурье от функции $\psi(t)$, тогда

$$\int_0^{\infty} \frac{|\hat{\psi}(\lambda)|^2}{\lambda} d\lambda < \infty.$$

На Рис. 26 показаны примеры вейвлетов, которые часто используются на практике.

Непрерывное вейвлет преобразование

Вейвлет $\psi(t)$, свойства которого были описаны выше, часто называют материнским или базовым вейвлетом. На основании материнского вейвлета строят семейство функций с помощью растяжения/сжатия и параллельного переноса. Это необходимо, чтобы исследовать различные области исходного сигнала и с различной степенью детальности.

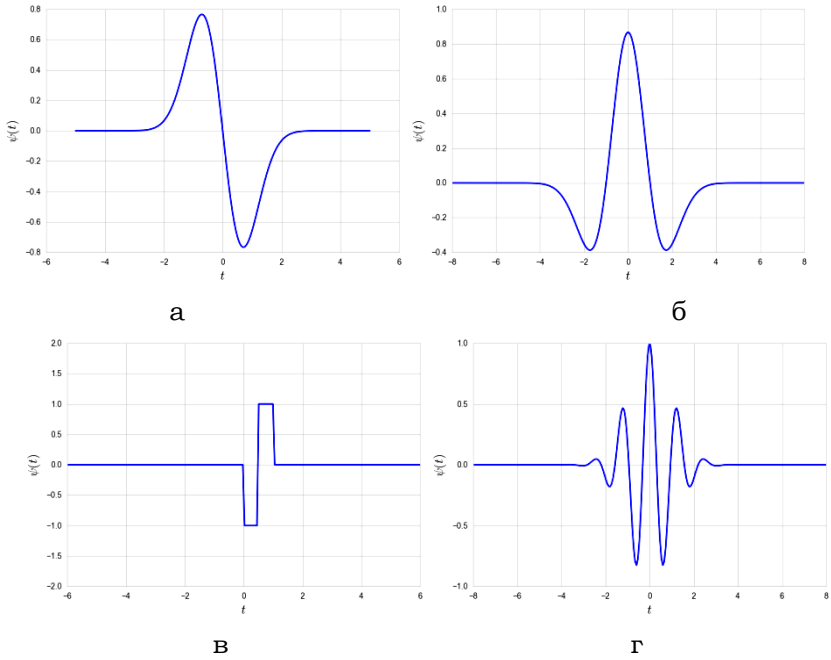


Рис. 26 – Примеры вейвлетов, которые часто используются в приложениях: (а) гауссова волна (первая производная гауссовой функции), (б) мексиканская шляпа, (в) вейвлет Хаара, (г) вейвлет Морле (действительная часть).

Введем параметры масштаба s (scale) и сдвига l (location), тогда преобразованная версия материнского вейвлета будет следующей

$$\psi_{s,l}(t) = \frac{1}{\sqrt{|s|}}\psi\left(\frac{t-l}{s}\right).$$

Непрерывным вейвлет преобразованием функции $x(t) \in L^2(\mathbb{R})$ называется выражение

$$W(s,l) = \frac{1}{\sqrt{|s|}} \int_{-\infty}^{\infty} x(t)\psi^*\left(\frac{t-l}{s}\right) dt = \int_{-\infty}^{\infty} x(t)\psi^*_{s,l}(t)\, dt,$$

где $l, s \in \mathbb{R}$, $s \neq 0$; ψ^* - функция комплексно сопряженная с ψ, величины $\{W(s,l)\}_{l,s\in\mathbb{R}}$ называются коэффициентами вейвлет преобразования.

Из формулы в определении непрерывного вейвлет преобразования непосредственно видно, что суть такого преобразование состоит в вычислении корреляционных коэффициентов специального вида. На Рис. 27 показано как на исходный ряд накладывается вейлет мексиканская шляпа и определяется корреляция между частью ряда и “шаблоном”, который представляет собой вейвлет. Затемненная область показывает вклад в значение вейвлет преобразования для данных сдвига и масштаба.

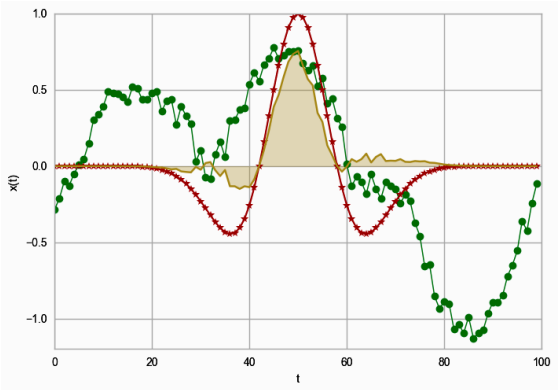


Рис. 27 – Иллюстрация к вычислению вейвлет преобразования как вычислению корреляции между исходным сигналом и функцией вейвлетом

Стоит заметить, что непрерывное вейвлет преобразования является обратимой операцией. Обратное вейвлет преобразование выполняется следующим образом:

$$x(t) = \frac{1}{C_g} \int_{-\infty}^{\infty} \int_{0}^{\infty} W_x(s,l)\psi_{s,l}(t) \frac{dsdl}{s^2}.$$

Проиллюстрируем результаты вейвлет преобразования на нескольких простых примерах (Рис. 15). Первый пример – сумма двух колебательных процессов. Под графиком сигнала показаны коэффициенты вейвлет преобразования, которые были получены при использовании вейвлета мексиканская шляпа. Вдоль горизонтальной оси изменяется время (параметр сдвига l), вдоль вертикальной оси – масштаб (параметр s). На графике вейвлет коэффициентов можно увидеть два периодических процесса. На Рис. 28 можно увидеть, как отражаются на значении вейвлет коэффициентов периодические процессы с разной амплитудой и частотой, а также отдельные пики в сигнале.

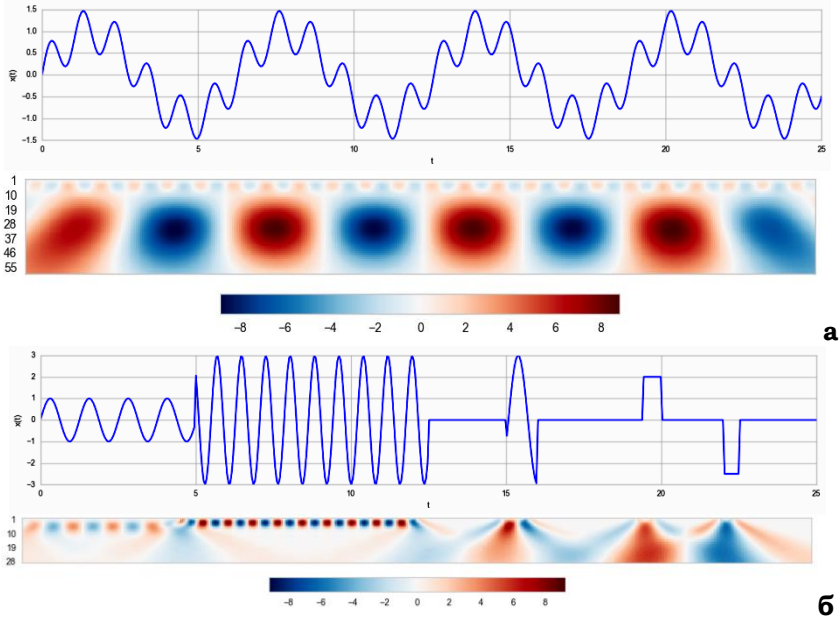


Рис. 28 – Примеры вейвлет преобразования. Сумма двух синусоид и ее вейвлет-преобразование (а). Функция, составленная из колебательных процессов разной частоты и амплитуды, а также отдельных пиков (б).

Рассмотрим коэффициенты вейвлет преобразования для временных рядов Т, К, Х. На Рис. 29 а, б показаны результаты вейвлет преобразования для ряда Т с использование вейвлета мексиканская шляпа (а) и гауссовой волны (б). На Рис. 29 в, г показаны результаты вейвлет преобразования для рядов К и Х с использованием вейвлета «мексиканская шляпа».

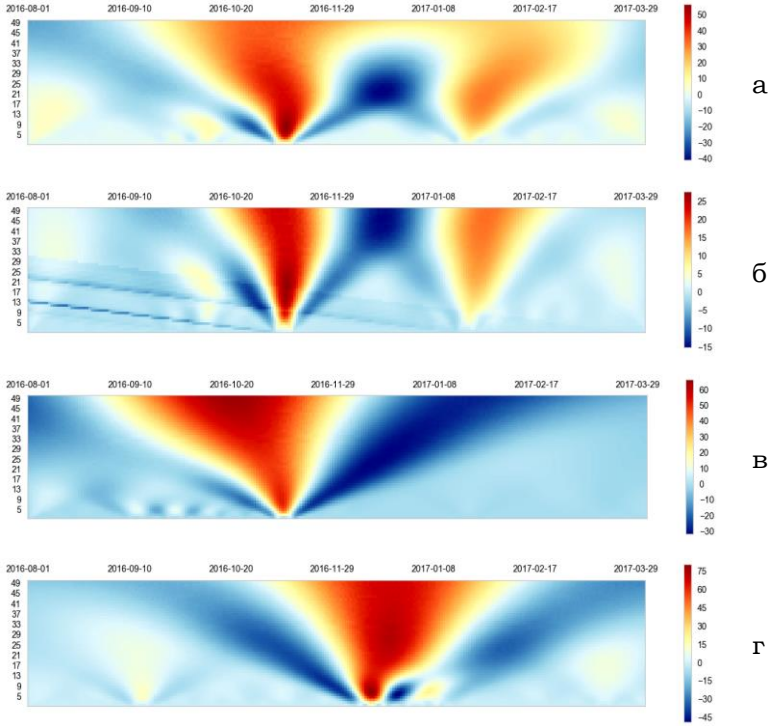


Рис. 29 – Вейвлет-преобразование для ряда Т с использование вейвлета мексиканская шляпа (а); действительная часть вейвлет-преобразования для ряда Т с использование вейвлета Морле (б); вейвлет преобразование для ряда К (в) и ряда Х (г) с использование вейвлета мексиканская шляпа

Еще раз заметим, что непрерывное вейлет преобразование, как и преобразование Фурье можно рассматривать в терминах корреляции. Преобразование Фурье – это корреляция

между исходным временным рядом и волной $\varphi(t) = e^{-i2\pi\nu t}$. Волна покрывает всю временную ось и характеризуется только частотой ν, поэтому преобразование Фурье зависит только от частоты. Вейвлет преобразование – это корреляция между исходным временным рядом и вейвлетом $\psi(t)$. Таким образом, вейвлет преобразование зависит от положения вейвлета на временной оси и его масштаба, которые определяются параметрами l и s соответсвенно.

Энергия сигнала

Полная энергия сигнала $x(t)$ по определению равна

$$E = \int_{-\infty}^{\infty} |x(t)|^2 dt.$$

Используя коэффициенты вейвлет преобразования можно определить энергию сигнала, которая соответствует определенному сдвигу и масштабу

$$E(s,l) = |W(s,l)|^2.$$

Значения $E(s,l)$ можно изобразить на графике, точно также как и коэффициенты вейвлет преобразования. Такой график обычно называют скейлограммой (scalogram). Также можно определить относительный вклад энергии, которая соответствует определенному масштабу, в полную энергию, или другими словами распределение энергии в зависимости от масштаба

$$E(s) = \frac{1}{C_g} \int_{-\infty}^{\infty} |W(s,l)|^2 dl.$$

Заметим, что на скейлограмме, в отличие от графика с коэффициентами вейвлет преобразования, все значения положительны, и наиболее ярко выделяются области с наибольшей энергией (Рис. 30).

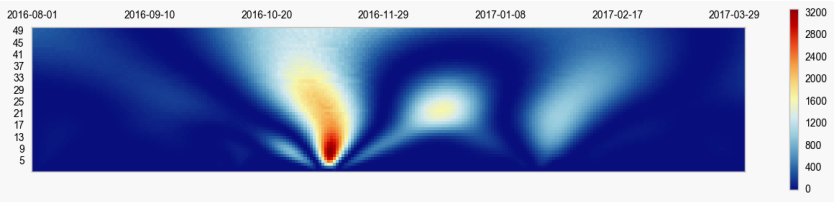


Рис. 30 – Скейлограмма для временного ряда T

Сравнение временных рядов с помощью вейвлет преобразования

Ранее был показан способ выявления зависимости между двумя временными рядами с помощью кросс-корреляции. Сейчас мы рассмотрим некоторые способы сравнения временных рядов с помощью коэффициентов вейвлет преобразования. Эти способы также можно применять для того, чтобы выявить некоторый тип отношения или взаимосвязи между временными рядами. Метрики сравнения коэффициентов вейвлет преобразования, а также примеры их применения к реальным практическим задачам, подробно описаны в [Addison 2017].

Рассмотрим два временных ряда x_t и y_t, и обозначим коэффициенты вейвлет преобразования этих рядов $W_x(s,l)$ и $W_y(s,l)$. Начнем с простейшего способа сравнения – возьмем разность модулей соответствующих коэффициентов

$$DiffMOD_{x,y}(s,l) = |W_x(s,l)| - \left|W_y(s,l)\right|.$$

На Рис. 31 показаны значения $DiffMOD_{x,y}(s,l)$ для двух пар рядов – сверху для рядов T и K, а снизу для рядов T и X. Таким простым способом можно выделить области в которых коэффициенты вейвлет преобразования схожи, а значит и в исходных временных рядах есть похожие участки.

Другим простым способом сравнения является отношение модулей коэффициентов вейвлет преобразования

$$RatioMOD_{x,y}(s,l) = \frac{|W_x(s,l)|}{\left|W_y(s,l)\right|}.$$

Такую метрику нужно использовать с осторожностью, так как $W_y(s,l)$ принимать нулевые или близкие к нулю значения.

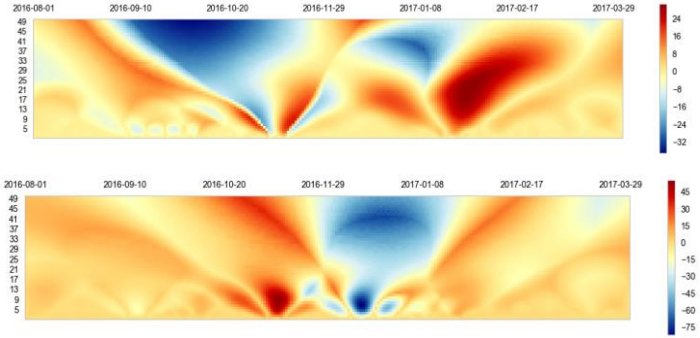


Рис. 31 – $\boldsymbol{DiffMOD_{x,y}(s,l)}$ для рядов T и K (а), T и X (б)

Дополнительную информацию можно получить, если использовать комплексный вейвлет (например, вейвлет Морле). Тогда кроме абсолютного значения вейвлет коэффициентов появляется фаза. Комплексный коэффициент всегда можно представить в виде

$$W(s,l) = |W(s,l)|e^{i\phi(s,l)},$$

где, как известно, модуль числа равен

$$|W(s,l)| = \sqrt{\mathrm{Re}\big(W(s,l)\big)^2 + \mathrm{Im}\big(W(s,l)\big)^2},$$

а фаза

$$\phi(s,l) = \tan^{-1}\left[\frac{\mathrm{Im}\big(W(s,l)\big)}{\mathrm{Re}\big(W(s,l)\big)}\right].$$

Следовательно, можно также сравнить фазы коэффициентов

$$\Delta\phi_{x,y}(s,l) = \phi_x(s,l) - \phi_y(s,l)$$

Кросс-вейвлет преобразование используется для выделения областей одинаковой энергии между сигналами в области преобразования, а также определения относительной фазы

$$CrWT_{x,y}(s,l) = W_x^*(s,l)W_y(s,l).$$

На рисунках обычно отображают значение $\left|CrWT_{x,y}(s,l)\right|$, по аналогии со скейлограммой. В таком случае, если временной

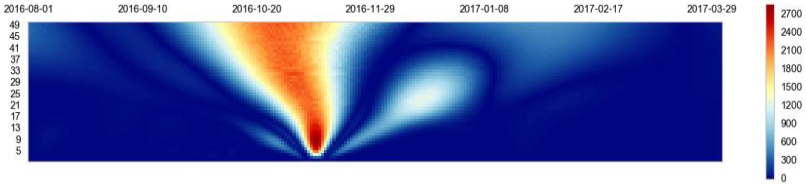
ряд x идентичен ряду y, то мы получим скейлограмму для ряда x.

Представляет особый интерес вычисление кросс-вейвлет преобразования в случае, когда используется комплексный вейвлет (например, вейвлет Морле). Тогда

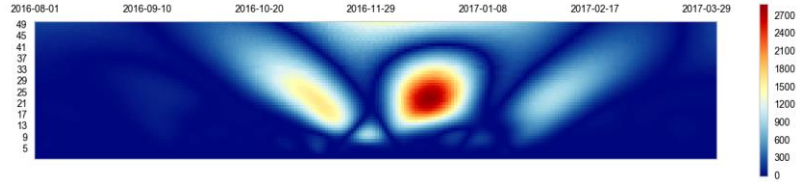
$$CrWT_{x,y}(s,l) = W_x^*(s,l)W_y(s,l) = |W_x(s,l)|e^{-i\phi_x(s,l)}\left|W_y(s,l)\right|e^{i\phi_y(s,l)} = \\ = |W_x(s,l)|\left|W_y(s,l)\right|e^{i\left(\phi_y(s,l)-\phi_x(s,l)\right)}.$$

Таким образом, вычисляя кросс-вейвлет преобразование можно извлечь значение разности фаз между коэффициентами вейвлет преобразования для двух врменных рядов.

На Рис. 32 показаны значения $CrWT_{x,y}(s,l)$ для радов Т и К. Для рядов Т и К выделяется область соответствующая пику интереса во время выборов, что означает, что это область высокой энергии для обоих рядов.



а



б

Рис. 32 – Кросс-вейвлет преобразование с использованием вейвлета мексиканская шляпа для рядов Т и К (а) и рядов Т и Х (б)

Если проинтегрировать коэффициенты вейвлет преобразования по времени, то получим вейвлет-кросс-корреляционную

меру (Wavelet Cross-Correlation Measure), которая зависит от масштаба

$$W_{x,y}(s) = \frac{\left|\int W_x^*(s,l)W_y(s,l)dl\right|}{\sqrt{\int|W_x(s,l)|^2dl\int\left|W_y(s,l)\right|^2dl}} = \frac{\left|\int CrWT_{x,y}(s,l)dl\right|}{\sqrt{\int|W_x(s,l)|^2dl\int\left|W_y(s,l)\right|^2dl}}.$$

Такая мера помогает обнаружить корреляцию между сигналами, которые содержат колебания с разной амплитудой или фазой, но, тем не менее, коррелированны между собой (Рис. 33).

Также можно расширить определение вейвлет кросс-корреляционной меры, если ввести зависимость от сдвига между рядами

$$W_{x,y}(s,k) = \frac{\left|\int W_x^*(s,l)W_y(s,l-k)dl\right|}{\sqrt{\int|W_x(s,l)|^2dl\int\left|W_y(s,l)\right|^2dl}}$$

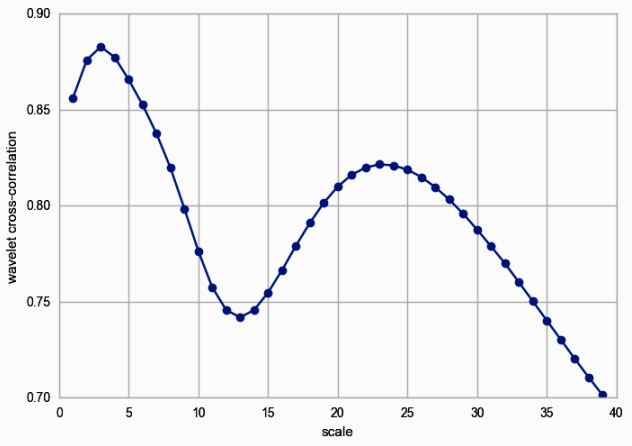


Рис. 33 – Зависимость вейвлет-кросс-корреляционной меры от масштаба для временных рядов T и K

Для исследования связи между компонентами временных рядов локально в плоскости преобразования, мы можем определить квадратичную оценку когерентности вейвлетов следующим образом:

$$WCH^2_{x,y}(s,l) = \frac{\left|\langle W_x^*(s,l) W_y(s,l)\rangle\right|^2}{\langle |W_x(s,l)|^2\rangle \langle \left|W_y(s,l)\right|^2\rangle},$$

где $\langle\cdot\rangle$ обозначает локальную операцию сглаживания, как во временной шкале, так и в масштабной, при этом сглаживание выполняется на компонентах преобразования.

Методы кросс-вейвлет анализа используют при исследовании свойств нескольких временных рядов, зависимых между собой нетривиальным образом. Например, в геофизике возникает задача выявления причинно-следственных связей или корреляции между метеорологическими или другими явлениями окружающей среды, которые происходят на большом расстоянии друг от друга. В [Maraun, 2004] анализируются особенности применения кросс-вейвлет преобразования и оценки когерентности вейвлетов для двумерных временных рядов такого типа. Также методы вейвлет анализа, в том числе и кросс-вейвлет преобразование, были использованы в [Adamowski, 2008] для изучения метеорологических временных рядов и данных об уровне потока реках. Из временных рядов двух типов выделяли компоненты, которые далее использовались в модели прогнозирования наводнений. В работе показано, что использование кросс-вейвлет анализа полезно в том случае, когда существует относительно стабильный сдвиг фазы между потоковым и метеорологическим временными рядами. С помощью кросс-вейвлет преобразования определялась разность фаз между потоковыми и метеорологическими данными, что улучшило качество модели прогнозирования наводнений.

Другим примером является [Labat, 2010], где кросс-вейвлет анализ проводился для климатических индексов и показателей сброса пресной воды в Африке. В этом случае, кросс-вейвлет преобразование и оценка когерентности использовались для визуализации и анализа периодических колебаний в данных длиной в 2-8 лет. В [Kelly, 2003] подтверждается, что методы на основе кросс-вейвлет анализа могут быть эффективным инструментом в поиске квазипериодичности временного ряда.

Методы кросс-вейвлет анализа также нашли применение в медицине. В [Li, 2007] описывается использование кросс-

вейвлет преобразования, оценки когерентности и некоторых других методов на основе вейвлетов для исследования динамики взаимодействия между колебаниями, генерируемыми двумя анатомически различными группами нейронов. Результаты исследования могут быть использованы для анализа и количественного определения временного взаимодействия между нейронными осцилляторами, а также для исследования механизмов эпилепсии.

В [Aguiar-Conraria, 2008] используются инструменты кросс-вейвлет анализа, чтобы показать, что связь между переменными денежной политики и макроэкономическими переменными со временем изменилась, причем эти изменения не являются однородными на разных частотах.

Данные, полученные с помощью кросс-вейвлет преобразования, также могут использоваться как исходные данные для алгоритмов классификации. В [Dey, 2010] коэффициенты кросс-вейвлет преобразования подавались на вход искусственной нейронной сети и классификатора Fuzzy.

Дискретное вейвлет преобразование и приближение функций с помощью ряда

Дискретное вейвлет преобразование определяется таким образом, чтобы можно было полностью восстановить исходный сигнал, используя бесконечные суммы дискретных вейвлет коэффициентов. Такой подход также приводит к быстрому вычислению вейвлет преобразования и его обратного [Addison, 2017].

Пусть $x(t)$ принадлежит пространству 2π −периодических квадратично интегрируемых функций. Тогда $x(t)$ можно представить в виде ряда Фурье

$$x(t) = \sum_{n=-\infty}^{\infty} c_n \, e^{int},$$

где коэффициенты c_n имеют вид:

$$c_n = \frac{1}{2\pi} \int_0^{2\pi} x(t) e^{-int} dt.$$

Набор функций $w_n(t) = e^{int}$ – это ортонормированный базис в пространстве $L^2(0, 2\pi)$, построенный с помощью масштабного преобразования $w_n(t) = w(nt)$ из базовой функции $w(t) = e^{int}$.

Пусть теперь $x \in L^2(\mathbb{R})$. Базисной функций в пространстве $L^2(\mathbb{R})$ должна быть функция, которая достаточно быстро убывает к 0 на $\pm\infty$. Поэтому, для построения базиса используются вейвлеты – хорошо локализованные солитоноподобные функции. Для того чтобы покрыть вейвлетами всю действительную ось, используют перенос вдоль оси. Для простоты можно использовать целые сдвиги k и аналоги синусоидальной частоты, как степени двойки $\psi_{jk} = \psi(2^j t - k)$. Вейвлет $\psi \in L^2(\mathbb{R})$ называется ортогональным, если семейство функций $\{\psi_{jk}\}$ образует ортонормированный базис в $L^2(\mathbb{R})$.

2.6.5. Корреляция с шаблоном

С помощью непрерывного вейвлет-преобразования выявляются участки исследуемого ряда, которые по форме наиболее похожи на вейвлет (Рис. 34). Идея состоит в том, чтобы сравнить части ряда с некоторым шаблоном на разных масштабах (Рис. 35). При этом вейвлет как функция должен обладать определенными математическими свойствами, в частности быстро убывать к нулю на бесконечности. В некоторых случаях полезно использовать шаблон, который не соответствует требованиям к вейвлету. Для этого вместо вейвлет-преобразования будем вычислять корреляцию между частью временного ряда и некоторым шаблоном p

$$\mathrm{C}(l,k) = \frac{\sum_{i=1}^{k}(x_{l+i} - \bar{x})\,(p_i - \bar{p})}{\sqrt{\sum_{i=1}^{k}(x_{l+i} - \bar{x})^2 \sum_{i=1}^{k}(p_i - \bar{p})^2}},$$

Полученный коэффициент $\mathrm{C}(l,k)$ зависит от значений $x_{l+1}, \ldots, x_{l+k}$. То есть параметр l отвечает сдвигу шаблона, а параметр k соответствует количеству точек в шаблоне и в рассматриваемом отрезке ряда. Параметр k в данном случае является аналогом масштаба s, который использовали при вейвлет-преобразовании.

Если при вычислении коэффициента вейвлет-преобразования всегда использовался весь временной ряд, то в данном случае для вычисления $C(l,k)$ используются k точек ряда и шаблон длины k.

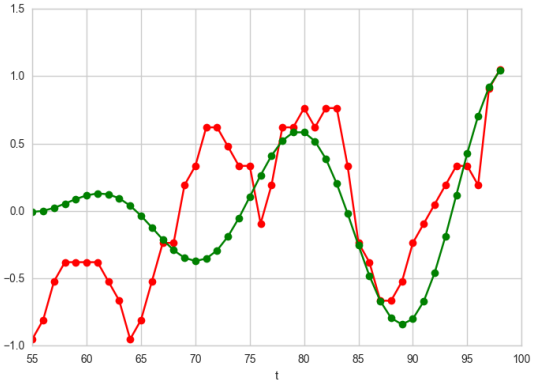


Рис. 34 – Отрезок временного ряда с наложенным шаблоном

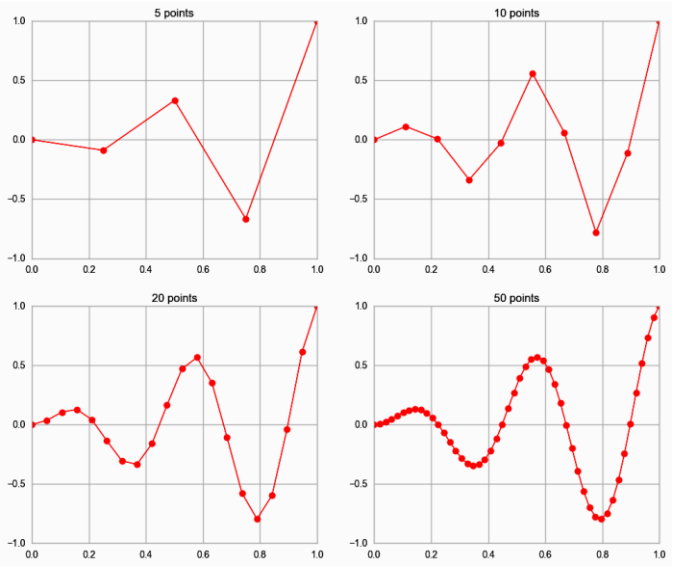


Рис. 35 – Шаблон «змея» с разным количеством точек

Полученные корреляционные коэффициенты $C(l,k)$ представим на графике, который похож на скейлограмму (Рис. 36).

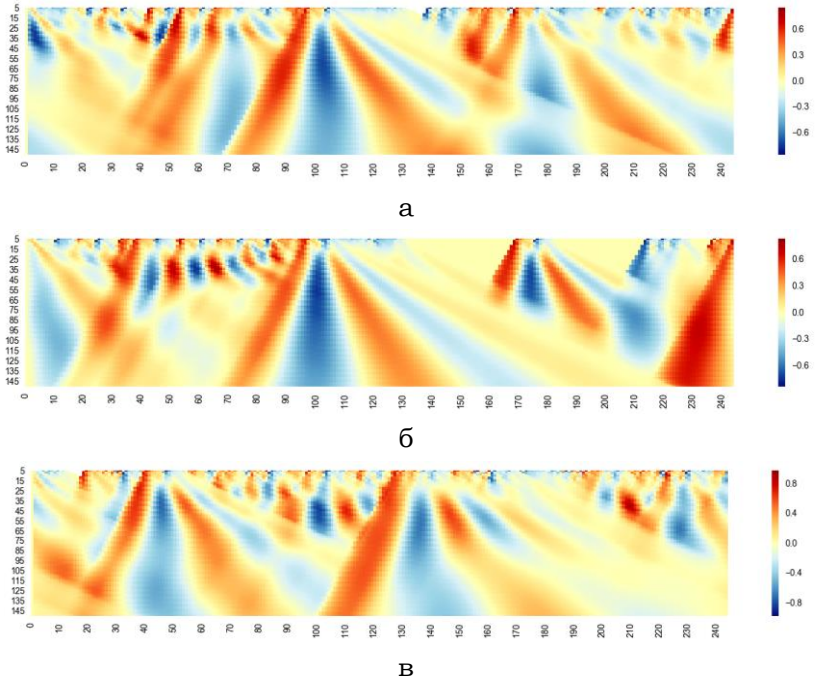


Рис. 36 – Корреляционные коэффициенты $\mathbf{C}(\boldsymbol{l},\boldsymbol{k})$ вычисленные для рядов T (а), K (б) и X (в) с использованием шаблона, показанного на Рис. 3.22

2.6.6. Фрактальный анализ

Термин фрактал ввел и популяризировал Бенуа Мандельброт. Чаще всего фракталами называют геометрические объекты, которые имеют сильно изрезанную форму и обладают свойством самоподобия.

Строго и общепринятого определения фрактала в данный момент не существует, хотя Бенуа Мальдеброт использовал несколько пробных определений. Одно из них, введенное в [Mandelbrot, 1982], звучит так:

Фракталом называется множество, размерность Хаусдорфа-Безиковича которого строго больше его топологической размерности.

Строгое определение размерности Хаусдорфа-Безиковича или фрактальной размерности будет введено позже. Суть такого определения сводится к тому, чтобы выделить класс сильно изрезанных объектов, для описания которых недостаточно топологической размерности. Например, существуют кривые, топологическая размерность которых всегда равна 1, но они изогнуты таким сложным образом, что заполнят плоскость или пространство. Так кривые Пеано, проходят через любую точку единичного квадрата. Другой пример – траектория броуновской частицы, которая не является гладкой ни в одной точке.

Первое определение, хотя и является строгим, но исключает многие физические фракталы, и поэтому не используется. Было предложено следующее определения фрактала:

Фракталом называется структура, состоящая из частей, которые в каком-то смысле подобны целому.

Второе определение подчеркивает, что отличительным признаком для фрактала является самоподобие. Приведем строгое определение самоподобного множества, которое используется в математике. Для этого потребуется несколько предварительных определений.

Пусть множество $E \subset \mathbb{R}^d$ замкнуто. Тогда отображение $S: E \to E$ называется отображением подобия (similarity) на E, если $\exists t: \ 0 < t < 1: \ |S(x) - S(y)| = t|x - y|, \forall x, y \in E$.

То есть отображение подобия S превращает множество E в геометрически подобное множество.

Рассмотрим набор отображение подобия $S_1, \dots, S_m$. Множество $F \subseteq E$ является инвариантным относительно преобразований S_i, если

$$F = \bigcup_{i=1}^{m} S_i(F).$$

Множество, которое является инвариантным относительно набора отображений подобия, называется самоподобным.

Определение самоподобного множества можно понять интуитивно. Действительно, по определению множество самоподобно, если его можно «собрать» из кусочков, которые подобны целому множеству. Тогда простейшим примером самоподобного множества будет отрезок $[0,1]$. Возьмем, например $S_1 = \frac{x}{2}$ и $S_2 = \frac{1}{2} + \frac{x}{2}$. Тогда $[0,1] = S_1([0,1]) + S_2([0,1])$.

Очевидно, что простого самоподобия не достаточно для того, чтобы назвать объект фракталом. В самом деле, не будем же мы считать фракталами отрезок прямой или листок бумаги в клеточку. Мы будем рассматривать фрактальные объекты, которые обладают свойством самоподобия, а также сложной структурой.

Приведем один базовый простейший пример фрактального множества, который будет удобно использовать для демонстрации основных идей в дальнейшем. Это множество Кантора или канторова пыль. Классический процесс построения канторового множества начинается с единичного отрезка $C_0 = [0,1]$. Удалим из C_0 среднюю треть, останется множество $C_1 = \left[0, \frac{1}{3}\right] \cup \left[\frac{2}{3}, 1\right]$. Множество C_1 состоит из двух отрезков, из каждого из которых теперь удалим среднюю треть, получим множество C_2. Продолжая повторять эту процедуру, получим последовательность множеств $\{C_i\}_{i=1}^{\infty}$. Множество Кантора – это пересечение

$$C = \bigcap_{i=1}^{\infty} C_i.$$

Заметим, что множество C самоподобно. Возьмем отображения подобия $S_1(x) = \frac{x}{3}$ и $S_2(x) = \frac{x}{3} + \frac{2}{3}$, тогда $C = S_1(C) \cup S_2(C)$. С другой стороны известно, что мера Лебега множества Кантора равна 0, точно также как и для точки, или любого счетного множества. Но очевидно, что структура множества Кантора намного более сложная, и здесь уже возникает идея, что для описания такого множества нужна специальная мера, которой и станет фрактальная размерность множества.

Фрактальная размерность

Вернемся к определению размерности Хаусдорфа-Безиковича. Для этого нам понадобится понятие покрытия множества.

Пусть U – непустое множество в $\mathbb{R}^d$. Диаметр множества Uпо определению равен

$$|U| = \sup\{|x - y| : x, y \in U\}.$$

Если $F \subset \bigcup_{i=1}^{\infty} U_i$ и $0 < |U_i| \leq \delta$ для любого i, то набор множеств $\{U_i\}$ называется δ −покрытием для множества U.

Пусть F – подмножество некоторого замкнутого множества в $\mathbb{R}^d$. Для произвольных $s \geq 0$ и $\delta > 0$ определим

$$\mathcal{H}_\delta^s(F) = \inf\left\{\sum_i |U_i|^s : \{U_i\} - \delta - \text{покрытие для } F\right\},$$

где инфинум берется по всем возможным δ −покрытиям множества F. По определению s-размерная мера Хаусдорфа

$$\mathcal{H}^s(F) = \lim_{\delta \to 0} \mathcal{H}_\delta^s(F).$$

Такой предел существует для любого множества $F \subset \mathbb{R}^d$, с той оговоркой, что часто он равен нулю или бесконечности.

Размерность Хаусдорфа-Безиковича множества F определяется как

$$D_H(F) = \inf\{s : \mathcal{H}^s(F) = 0\} = \sup\{s : \mathcal{H}^s(F) = \infty\}.$$

Или, что то-же самое

$$\mathcal{H}^s(F) = \begin{cases} 0, s < D_H(F); \\ \infty, s > D_H(F). \end{cases}$$

Для примера вычислим размерность множества Кантора C. Выше была описана процедура построения, и согласно с ней на n-м шаге имеется 2^n отрезков длины $1/3^n$ каждый и далее множество только уменьшается. Поэтому в качестве диаметра покрытия δ можно взять величину $1/3^n$ и использовать 2^n множеств в покрытии. По определению

$$\mathcal{H}^s(C) = \lim_{\delta \to 0} \mathcal{H}_\delta^s(C),$$

и теперь можно перейти от предела по $\delta \to 0$, к пределу $n \to \infty$

$$\lim_{\delta\to 0}\mathcal{H}_{\delta}^{s}(C)=\lim_{n\to\infty}\sum_{i=1}^{2^n}\left(\frac{1}{3^n}\right)^{s}.$$

Остается определить значение такого предела

$$\lim_{n\to\infty}\sum_{i=1}^{2^n}\left(\frac{1}{3^n}\right)^{s}=\lim_{n\to\infty}\left(\frac{2}{3^s}\right)^{n}=\begin{cases}0,\frac{2}{3^s}<1,\\ 1,\frac{2}{3^s}=1,\\ \infty,\frac{2}{3^s}>1.\end{cases}$$

Следовательно, предел $\mathcal{H}_{\delta}^{s}(C)$ при $\delta\to 0$ не равен нулю или бесконечности при $\frac{2}{3^s}=1$, поэтому

$$D_H(C)=\frac{\ln 2}{\ln 3}\approx 0{,}63.$$

Статистически самоподобные процессы и показатель Херста

Многие объекты в окружающем нас мире статистически самоподобны (классический пример – береговые линии), это означает, что части таких объектов имеют одинаковые статистические характеристики при изменении масштаба. При изучении эволюции информационных потоков, структуры массивов документов в Интернет и исследовании процессов в информационном пространстве часто возникают самоподобные структуры, и в частности временные ряды.

Дадим определение самоподобного процесса.

Действительнозначный процесс $\{x(t), t\in\mathbb{R}\}$ является самоподобным с показателем Херста $H>0$, если для всех $\alpha>0$ конечномерные распределения $\{x(\alpha t), t\in\mathbb{R}\}$ идентичны конечномерным распределениям $\{\alpha^{H}x(t), t\in\mathbb{R}\}$, что можно кратко записать

$$\{x(\alpha t), t\in\mathbb{R}\}=^{d}\{\alpha^{H}x(t), t\in\mathbb{R}\}.$$

То есть, по определению, для самоподобного процесса изменение временного масштаба эквивалентно изменению масштабу значений процесса. Это означает, что реализации такого процессы выглядят одинаково на разных масштабах. При этом, естественно, что процесс не является точной копи-

ей себя на разных масштабах, сохраняются только статистические свойства.

Показатель Херста представляет собой меру персистентности — склонности процесса к трендам. Значение H = 0.5 соответствует некоррелированному поведению значений ряда, как у броуновского движения. Значения в диапазоне 0.5 <H < 1 означают, что направленная в определенную сторону динамика процесса в прошлом, вероятнее всего, повлечет продолжение движения в том же направлении. Если же H > 0.5, то прогнозируется, что процесс изменит направленность.

Опишем некоторые свойства самоподобных процессов, которые важны для приложений. Во-первых, у таких процессов автоковариационная функция гиперболически затухает, и имеет вид

$$\rho_k \approx k^{(2H-2)}L(t) \text{ при } k \to \infty,$$

где $L(t)$ – медленно меняющаяся на бесконечности функция, то есть такая, что

$$\forall x > 0: \lim_{t\to\infty} \frac{L(tx)}{L(t)} = 1.$$

Следовательно, для самоподобных процессов ряд из ковариационных коэффициентов расходится

$$\sum_{k=1}^{\infty} \rho_k = \infty.$$

Такая бесконечная сумма говорит о долговременной зависимости в ряде.

Во-вторых, дисперсия выборочного среднего убывает медленнее, чем величина, обратная размеру выборки

$$\sigma^2\left(x_t^{(m)}\right) \sim m^{2H-2},$$

где последовательность $\left\{x_t^{(m)}\right\}$ получили, разбив исходную последовательность $\{x_t\}$ на непересекающиеся блоки длины m и взяв среднее в каждом из блоков.

Методы оценивания показателя Херста

Метод оценивания показателя Херста, предложенный им самим, называется методом нормированного размаха или R/S анализом. Для временного ряда $\{x_t\}_{t=1}^{T}$ стандартное отклонение S определяется по формуле

$$S = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(x_t - \bar{x})^2}, \quad \text{где } \bar{x} = \frac{1}{T}\sum_{t=1}^{T} x_t,$$

а величина размаха ряда

$$R = \max_{1\le t\le T} x^{(t)} - \min_{1\le t\le T} x^{(t)}, \text{где } x^{(t)} = \sum_{i=1}^{t}(x_i - \bar{x}).$$

Отношение R/S и есть нормированным размахом. Херст обнаружил, что для многих наблюдаемых временных рядов нормированный размах хорошо описывается эмпирическим соотношением

$$\frac{R}{S} = \left(\frac{T}{2}\right)^{H}.$$

Значения показателя Херста можно оценить, если вычислить значения статистики R/S в зависимости от T и построить график такой зависимости в двойной логарифмической шкале. Оценкой показателя Хёрста будет оценка наклона прямой, которая наилучшим образом аппроксимирует зависимость $\log R/S$ от $\log T$.

Используем метод R/S для вычисления показателя Хёрста для рядов T, K и X. На Рис. 37 показаны результаты оценивания для рядов T и K. Полученные значения показателя Хёрста – 0.62 и 0.68 соответственно – свидетельствуют о склонности данных процессов к трендам, хотя и не очень высокую.

В случае ряда K на Рис. 37 видно, что зависимость $\log R/S$ от $\log t$ плохо аппроксимируется линейной зависимостью, так как график имеет сильный излом. Если построить зависимость показателя Хёрста от времени (Рис. 38), то можно определить момент времени, начиная с которого значение показателя начинает убывать. Отметив этот момент времени на

графике временного ряда X, можно увидеть, что это момент резкого возрастания значений ряда, до которого значения ряда имели значительно меньшую дисперсию.

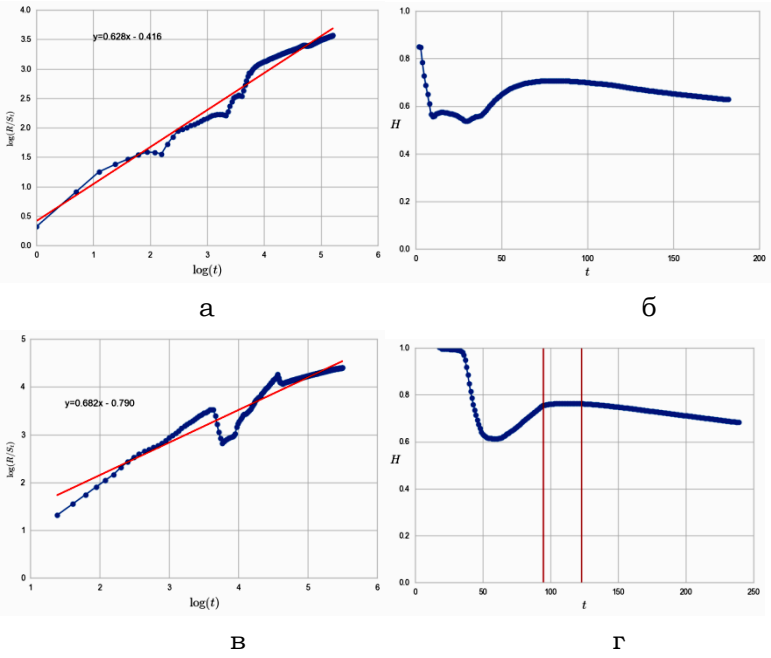
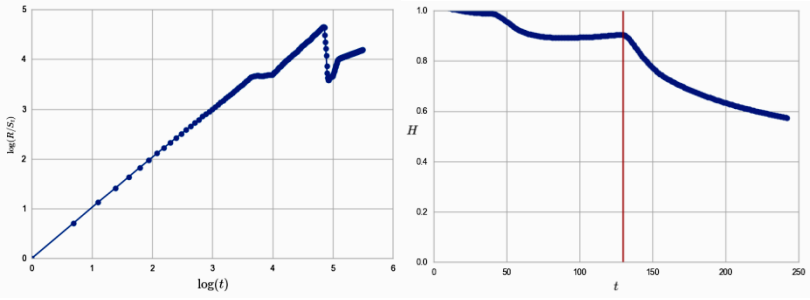
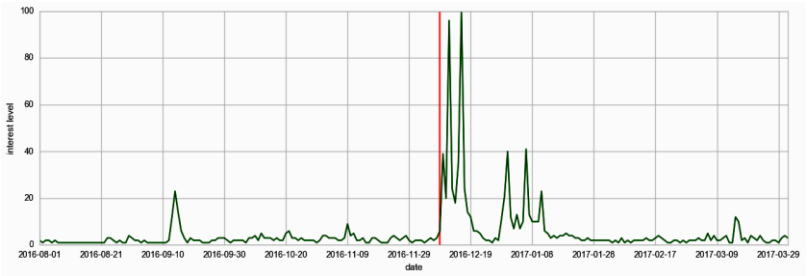


Рис. 37 – Оценивание показателя Херста для рядов T и K. Зависимость статистики $\boldsymbol{R/S}$ для ряда T (а) и ряда K (в) от времени в логарифмической шкале. Зависимость показателя Херста от времени для ряда T (б) и ряда K (г)

Поведение ряда X, начиная с начала декабря 2016 (начало наибольшего пика в значениях ряда) можно рассмотреть отдельно. Оценка показателя Хёрста для второго участка ряда показана на Рис. 39 При этом стоит учесть, что в данном примере временной ряд становится слишком коротким, так как для R/S анализа используют ряды с не менее чем 200 элементами. Тем не менее, резкие изменения в зависимости показателя Хёрста от времени, которые имеют вид «ступеньки», свидетельствуют о том, что исследуемый процесс состоит из различных процессов, которые имеет смысл рассмотреть отдельно.



а б

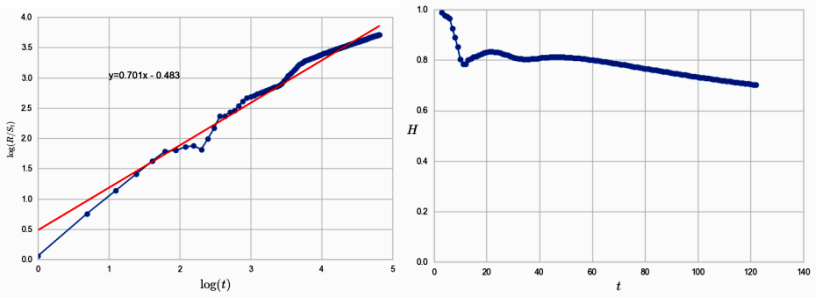


в

Рис. 38 – Оценивание показателя Херста для ряда X. Зависимость статистики $\boldsymbol{R/S}$ от времени в логарифмической шкале (а). Зависимость показателя Хёрста от времени (б). Вертикальной линией отмечен момент времени, после которого значение показателя Хёрста начинает убывать. Временной ряд X с отмеченным моментом времени, после которого происходит изменение показателя Хёрста (в)

ΔL-метод

Скейлограммы, полученные с помощью непрерывного вейвлет-преобразования, используют для визуализации особенностей временного ряда. В [Ландэ, 2009] предложен другой метод визуализации, который также помогает выявить тренды, периодичности и локальные особенности ряда. Предложенный подход значительно проще в реализации, чем вейвлет анализ.



а б

Рис. 39 – Оценивание показателя Херста для второй части ряда X. Зависимость статистики R/S от времени в логарифмической шкале (а). Зависимость показателя Хёрста от времени (б).

Метод, который авторы назвали Δ-метод, базируется на методе DFA (Detrended Fluctuation Analysis), который также будет рассмотрен ниже. Суть подхода состоит в определении и отображении абсолютного отклонения точек ряда накопленных значений от соответствующих значений линейной аппроксимации.

Опишем Δ-метод более подробно. Для начала зафиксируем некоторую ширину окна s (масштаб на котором рассматривается ряд). Рассмотрим точку x_l и выберем для нее окно ширины s так, чтобы точка l была в центре этого окна (или смещена на 1, если s чётное). Построим линейную аппроксимацию по точкам окна и обозначим $L_{l,j,s}$ значение локальной аппроксимации в точке j для отрезка с центром в l. Далее вычислим абсолютное отклонение x_j (Рис. 40) от линии аппроксимации $\Delta_{l,j,s} = \left|x_j - L_{t,j,s}\right|$.

Метод предполагает вычисление значений $\Delta_{l,j,s}$ для всех точек $l = 1, \dots, T$ и окон шириной $s = 1, \dots, [T/4\,]$. Для фиксированной ширины окна вычисляется среднее квадратичное отклонение

$$E(l,s) = \sqrt{\frac{1}{s}\sum_{j}\left|x_j - L_{l,j,s}\right|^2} = \sqrt{\frac{1}{s}\sum_{j}{\Delta_{l,j,s}}^2}.$$

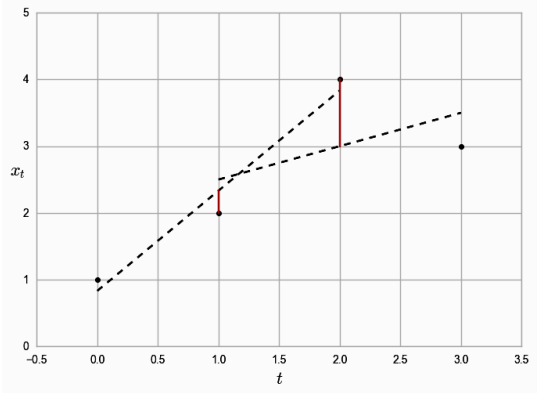


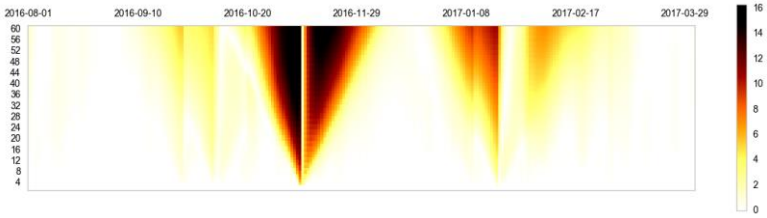
Рис. 40 – Четыре точки временного ряда с линейной аппроксимацией для двух окон с шириной три. Также показано отклонение $\boldsymbol{\Delta_{t,j,s}}$ центральной точки окна от соответствующей линейной аппроксимации.

Далее полученные значения демонстрируются на диаграмме, похожей на скейлограмму. Примеры таких диаграмм показаны на Рис. 41.

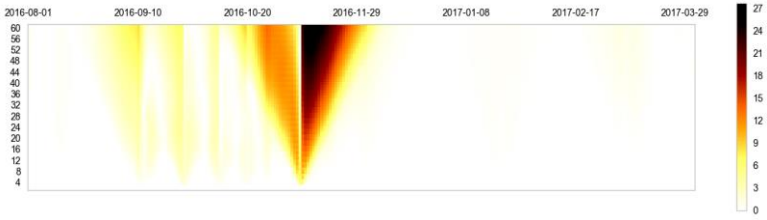
Предложенный метод визуализации абсолютных отклонений ΔL, как и метод вейвлет-преобразований, позволяет выявлять единичные и нерегулярные «всплески», резкие изменения значений количественных показателей в разные периоды времени, а также гармонические составляющие в ряде.

2.6.7. Мультифрактальный анализ

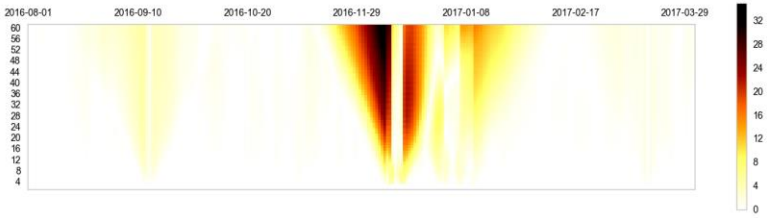
Для описания самоподобных объектов, которые возникают в природе, часто не хватает одной фрактальной размерности, так как во многих случаях такие объекты не являются однородными. Наиболее общее описание природы таких объектов дает теория мультифракталов, согласно которой объект характеризуется бесконечной иерархией размерностей, что позволяет отличить однородные объекты от неоднородных.



а



б



в

Рис. 41 – Коэффициенты полученные с помощью ΔL-метода для рядов Т (а), К (б) и Х (в).

Мультифрактальное множество (сигнал) можно понимать как некое объединение различных однородных фрактальных подмножеств (сигналов), каждое из которых имеет собственное значение фрактальной размерности. Значения таких фрактальных размерностей отображаются в мультифрактальном спектре, формальное определение которого будет введено позже. Важно, что мультифрактальный спектр может использоваться как мера подобия. Такой подход может использоваться, например, для формирования репрезентативных выборок из массивов документов, как дополнение традиционных методов, базирующихся на выявлении содержательного подобия документов. Практические приложения та-

кого подхода: предъявление пользователю обозримых результатов поиска, отражающих весь спектр документального массива или выделение подмножеств документов для дальнейших исследований [Ландэ, 2009a: Ландэ, 2009b].

Для того, чтобы детально рассмотреть идею мультифрактального множества и мультифрактального спектра понадобится несколько дополнительных понятий и определений. Будем использовать в качестве примера обобщенное множество Кантора, а также меру на этом множестве. Мультифрактальный спектр в этом примере появляется достаточно естественным и простым образом, поэтому помогает получить интуитивное понимание. Также для анализа мультифрактальности ключевое значение имеет показатель Гёльдера, определение которого будет введено чуть позже.

Начнем с построение обобщенного множества Кантора и определения на нем меры. Классическое множество Кантора можно обобщить несколькими способами. Например, можно использовать функцию сжатия $F_i(x) = rx + (1-r)i$, где $i = 0{,}1$ и $r \in \left(0, \frac{1}{2}\right)$, вместо $F_i(x) = \frac{1}{3}x + \frac{2}{3}i$. Обозначим такое множество $C(r)$. На обобщенном множестве Кантора можно вести равномерно распределенную меру, а можно определить p −меру с такой же функцией сжатия, но вероятностями p и $1-p$, где $p \in [0,1]$.

Более интересным обобщением является множество Кантора с переменными коэффициентами разделения. Пусть есть последовательность $\{r_j\}$, $r_j \in \left(0, \frac{1}{2}\right)$. Множество будем строить с помощью следующей итерационной процедуры. Пусть $C_0 = [0,1]$. Удалим из середины отрезка $[0,1]$ открытый отрезок длины $1-2r_1$. Останется два замкнутых отрезка длины r_1. Объединение этих двух отрезков обозначим C_1. На $j-$ой итерации множество C_j будет состоять из объединения 2^j замкнутых отрезков длины $r_1 \cdot r_2 \cdot \ldots \cdot r_j$. Таким образом, получим множество

$$C(\{r_j\}) = \bigcap_{j=1}^{\infty} C_j.$$

Замкнутые отрезки, которые появляются при итерационной конструкции множества Кантора, можно закодировать с помощью конечных слов из алфавита $\{0,1\}$. На первом шаге получаем левый отрезок I_0 и правый отрезок I_1. На $n-$ом шаге название отрезка w имеет длину n и на следующем шаге отрезок I_w будет разделен на отрезки I_{w0} и I_{w1}.

На множестве $C(\{r_j\})$ можно ввести $p-$меру, которая будет обладать следующим свойством

$$\mu(I_{w0}) = p\mu(I_w), \qquad \mu(I_{w1}) = (1-p)\mu(I_w).$$

Еще более обобщенную меру на $C(\{r_j\})$ получим, если использовать последовательность весов $\{p_j\}, p_j \in [0,1]$ и определить меру с помощью следующего правила $\mu(I_{w_1 w_2 \dots w_n}) = p_{w_1 1} \cdot p_{w_2 2} \cdot \dots \cdot p_{w_n n}$, где $p_{0j} = p_j$, а $p_{1j} = 1 - p_j$.

Для того чтобы сделать некоторые выводы о структуре меры понадобится показатель Гёльдера. Сначала рассмотрим определение и смысл показателя Гёльдера для функций и мер, а потом применим их к построенному только что обобщенному множеству Кантора и мере на нём.

Показатель Гёльдера и мультифрактальный анализ для мер

Характеристикой гладкости функции является показатель Гёльдера, который содержит информацию о поведении функции в окрестности точки. Чем меньше значение показателя Гельдера, тем менее гладкой является функция.

Пусть x – ограниченная функция на $\mathbb{R}$ и$t_0 \in \mathbb{R}$, тогда локальный показатель Гёльдера функции x в точке t_0 определяется как

$$h_x(t_0) = \sup_{\Delta t \to 0} \{\alpha \geq 0 : |x(t+\Delta t) - x(t)| = O(\Delta t^{\alpha})\}.$$

Другими словами, локальный показатель Гёльдера характеризует поведение функции в окрестности точки следующим образом

$$|x(t+\Delta t) - x(t)| \sim \Delta t^{h_x(t)}.$$

Последнее соотношение стоит сравнить с похожим соотношением для монофрактальных процессов

$$|x(t+\Delta t)-x(t)|\sim\Delta t^{H},$$

где H – это показатель Херста. То есть для мультифрактальных процессов локальный показатель Гельдера $h_x(t)$ по сути является «локальным показателем Херста», который может меняться в зависимости от t.

Для того чтобы измерить регулярность меры в окрестности точки также вводят показатель Гельдера.

Показатель Гёльдера или локальная размерность меры μ на $\mathbb{R}$ определяется следующим образом

$$\mathrm{h}_{\mu}(\mathrm{x})=\lim_{\mathrm{r}\to 0^{+}}\frac{\log[\mu(\mathrm{B}(\mathrm{x},\mathrm{r}))]}{\log \mathrm{r}},$$

где $\mathrm{B}(\mathrm{x},\mathrm{r})$ – шар с центром в точке x и радиусом r.

Вернемся к примерам с множеством Кантора. Для равномерной распределенной меры на классическом множестве Кантора:

$$\mathrm{h}_{\mu}(x)=\frac{\log 2}{\log 3}, x\in \mathrm{C}.$$

Если же мы рассмотрим множество $\mathrm{C}(\{\mathrm{r_i}\})$ *с мерой* μ, определенной с помощью весов p и $1-p$, то локальная размерность меры будет отличаться в разных точках. Например,

$$\mathrm{h}_{\mu}(0)=\lim_{\mathrm{j}\to\infty}\frac{\log \mathrm{p}^{\mathrm{j}}}{\log\prod_{\mathrm{i}=1}^{\mathrm{j}}\mathrm{r_i}}=\frac{\log \mathrm{p}}{\lim_{\mathrm{j}\to\infty}\frac{1}{\mathrm{j}}\sum_{\mathrm{i}=1}^{\mathrm{j}}\log \mathrm{r_j}}=\frac{\log \mathrm{p}}{\log \mathrm{r}_0}, \mathrm{h}_{\mu}(1)=\frac{\log(1-\mathrm{p})}{\log \mathrm{r}_0},$$

где ввели обозначение $\lim_{\mathrm{j}\to\infty}\frac{1}{\mathrm{j}}\sum_{\mathrm{i}=1}^{\mathrm{j}}\log \mathrm{r_j}=\log \mathrm{r}_0$. Если предположить, что $\sup_{\mathrm{i}} \mathrm{r_i}<1/2$, и без ограничения общности считать, что $\mathrm{p}<1-\mathrm{p}$, то можно показать, что

$$\mathrm{h}_{\mu}(\mathrm{x})\in\left[\frac{\log \mathrm{p}}{\log \mathrm{r}_0},\frac{\log(1-\mathrm{p})}{\log \mathrm{r}_0}\right], \mathrm{x}\in \mathrm{C}(\{\mathrm{r_i}\}).$$

(доказательство [Aldroubi 2016]). Таким образом, для $\mathrm{p}-$Канторовой меры μ известны возможные значения локальной размерности. Для того чтобы описать некоторые свойства

меры μ нужно рассмотреть множества уровня значений локальной размерности, а именно множества вида

$$E_h = \{x \in \mathbb{R}: h_\mu(x) = h\}.$$

Далее можно сравнить размеры множеств E_h при разных значениях h. Во многих практически важных случаях для сравнения таких множеств нужно будет использовать фрактальную размерность. Таким образом, приходим к определению мультифрактального спектра.

Мультифрактальным спектром меры μ на $\mathbb{R}$ называется отображение $d_\mu(h) = D_H(E_h)$.

То есть с помощью мультифрактального спектра отображается, какие значения показателя Гёльдера присутствую в неоднородном объекте (мере, множестве, сигнале), и в каком соотношении между собой. Каждому значению показателя Гёльдера соответствует фрактальная размерность множества точек, в которых значение показателя Гёльдера равно данному (Рис. 42).

Для p –Канторовой меры μ можно показать, что (доказательство [Aldroubi 2016])

$$d_\mu(h) = \inf_{q\in\mathbb{R}} \left(qh - \frac{\log(p^q + (1-p)^q)}{\log r_0} \right).$$

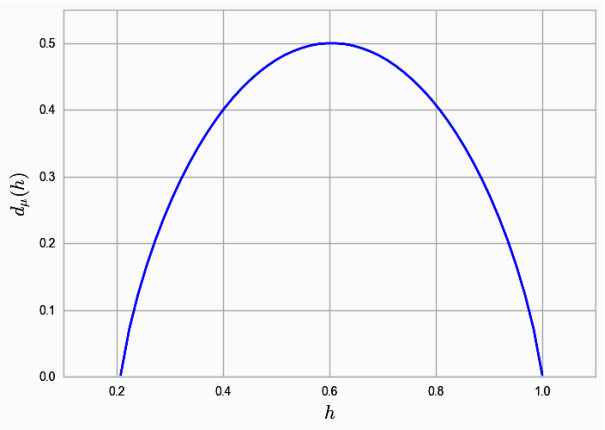


Рис. 42 – Мультифрактальный спектр для p-Канторовой меры.

Подход к оцениванию мультифрактального спектра

Выше был описан теоретический подход к определению мультифрактального спектра для меры. Для практических целей прямое вычисление показателя Гёльдера в каждой точке и вычисление фрактальных размерностей множеств уровня данного показателя не осуществимо. Ключ к практическому подходу дает следующее определение структурной функции (structure function) меры μ

$$Z(q,j)=\frac{1}{2^j}\sum_{\mathrm{i}}\mu\left(B\left(\frac{\mathrm{i}}{2^j},\frac{1}{2^{j+1}}\right)\right)^q,$$

где сумма берется только по таким отрезкам, где мера не равна 0. Также определим масштабную функцию (scaling function)

$$\tau(q)=\lim_{j\to\ +\infty}\inf\left(\frac{\log\left(Z_\mu(q,j)\right)}{\log 2^{-j}}\right).$$

Известно, что для того, чтобы покрыть множество $\mathrm{E_h}$ нужно приблизительно $2^{-d_\mu(h)j}$ шаров и из определения $h_\mu(x)=\lim_{\mathrm{j}\to\infty}\frac{\log[\mu(B(\mathrm{x},2^{-\mathrm{j}}))]}{\log 2^{-\mathrm{j}}}$ следует, что $\mu\left(B(\mathrm{x},2^{-\mathrm{j}})\right)\sim 2^{-\mathrm{h_\mu(x)j}}$, поэтому масштабную функцию можно оценивать следующим образом

$$Z(\mathrm{q,j})\sim 2^{-\mathrm{j}}\sum 2^{-\mathrm{h_\mu qj}}\sim 2^{-\mathrm{j}}2^{-\mathrm{d_\mu(h)j}}2^{-\mathrm{hqj}}=2^{-\left(1+\mathrm{d_\mu(h)+hq}\right)\mathrm{j}}.$$

С другой стороны, масштабную функцию определили таким образом, что $Z(\mathrm{q,j})\sim 2^{-\tau(\mathrm{q})\mathrm{j}}$, поэтому

$$2^{-\left(1+\mathrm{d_\mu(h)+hq}\right)\mathrm{j}}=2^{-\tau(\mathrm{q})\mathrm{j}}.$$

И при $\mathrm{j}\to\infty$

$$d_\mu(h)=\inf_{q\in\mathbb{R}}(1-\tau(q)+hq).$$

Таким образом, получили выражения для мультифрактального спектра через масштабную функцию. Данный подход позволяет численно определять мультифрактальный спектр для временных рядов. Сначала определяется структурная функция, с помощью нее – масштабная функция, и далее через преобразование Лежандра происходит переход к мультифрактальному спектру.

Мультифрактальные процессы

Стохастический процесс называется мультифрактальным, если он обладает стационарными приращениями и удовлетворяет равенству

$$\mathbb{E}[|x(t)|^q] = c(q)t^{\tau(q)+1},$$

для некоторого положительного q, где $\tau(q)$ – масштабная функция.

Если масштабная функция $\tau(q)$ линейно зависит от q, то процесс называется монофрактальным. Если процесс $x(t)$ самоподобный с показателем Херста H, то $\tau(q) = Hq - 1$.

Метод DFA и его применение к оцениванию мультифрактального спектра

В [Peng, 1994] предложен метод Detrended Fluctuation Analysis (DFA) для определения длительных корреляций в зашумленных и нестационарных временных рядах. Ключевая особенность метода DFA состоит в том, что он основан на теории случайных блужданий. Временной ряд не анализируется в исходном виде, вместо этого выполняется центрирование ряда и переход к накопленным сумам

$$y_t = \sum_{k=1}^{t} x_k.$$

В таком случае можно рассматривать y_t как положение случайного блуждания после t шагов. Далее метод DFA предполагает анализ среднеквадратического отклонения значений ряда от тренда на различных непересекающихся кусках ряда.

Для метода DFA было предложено множество модификаций, а также вариантов применения для различных практических задач. Обзор таких методов приводится, например, в [Kantelhardt 2009]. Важным шагом стала разработка подхода к численному оцениванию мультифрактального спектра на основе метода DFA. Такой метод называется Multifractal Detrended Fluctuation Analysis (MF-DFA) и был предложен в [Kantelhardt 2002]. Эффективность метода MF-DFA была проанализирована для различных модельных временных рядов (броуновское движение, дробное броуновское движение, би-

номиальные каскады) [Oswiecimka 2012]. Также метод активно используют для анализа реальных временных рядов, часто экономических [Suarez-Garcia 2013].

Подробное пошаговое описание алгоритма MF-DFA приведено в [Thompson 2016]. Опишем все эти шаги.

Шаг 1. Приведение временного ряда к агрегированному виду. Различают агрегированные и дисагрегированные наборы данных. Пример агрегированных данных – количество новых сообщений в сети Интернет по некоторой теме в день. Соответственные дисагрегированные данные – приращение количества сообщений по сравнению с предыдущим днем. Если исходный временной ряд $\{z_t\}_{t=1}^{T+1}$ агрегированный, то нужно перейти к дисагрегированному $\{y_t = z_{t+1} - z_t\}_{t=1}^{T}$. Временной ряд, который будет использоваться в алгоритме, при центрировании и подсчете накопленных сумм – следующий:

$$x_t = \sum_{k=1}^{t} (y_k - \bar{y}), \qquad t = 1, \dots, T.$$

Данный этап обработки ряда необходим для корректной работы метода, так как он основан на теории случайных блужданий.

Шаг 2. Зададим множество $\mathcal{S} = \{3, [N/4]\}$. Для каждого значения $s \in \mathcal{S}$ разделим временной ряд $\{x_k\}_{k=1}^{N}$ на $N_s = \left\lfloor \frac{N}{s} \right\rfloor$ непересекающиеся части длины s. Если N не делится нацело на s, то нужно повторить процедуру, начиная с другой стороны временного ряда, и в результате получится $2N_s$ частей.

Шаг 3. Для каждого значения $j = 1, \dots, N_s$ $j-$ая часть временного ряда состоит из значений $\{x_{(j-1)s+i}\}_{i=1}^{s}$, и аналогично при $j = N_s + 1, \dots, 2N_s$ – $\{x_{N-(j-N_s)s+i}\}_{i=1}^{s}$. Для каждой $j-$ой части временного ряда нужно определить тренд $X_j(i)$. Во многих случаях достаточно использовать линейную аппроксимацию значений ряда, полученную с помощью метода наименьших квадратов. Для достаточно длинных рядов используют полиномиальную аппроксимацию степени m (далее будет приведен алгоритм подбора оптимального параметра

m). Теперь вычислим среднеквадратичное отклонение части временного ряда от тренда

$$F^2(j,s) = \frac{1}{s}\sum_{i=1}^{s}\left[x_{(j-1)s+i} - X_j(i)\right]^2, \qquad \text{при } j = 1,\dots,N_s,$$

и аналогично при $j = N_s + 1,\dots,2N_s$.

Шаг 4. Обозначим Q множество значений порядков моментов. Множество Q должно содержать 0, положительные и отрицательные значения. Обычно выбирают симметричное относительно 0 множество. Для фиксированных значений $s \in S$ и $q \in Q$ вычислим норму l_q для вектора, состоящего из оцененных дисперсий $\{F^2(j,s): j = 1,\dots,2N_s\}$

$$F_q(s) = \left(\frac{1}{2N_s}\sum_{j=1}^{2N_s}[F^2(j,s)]^{q/2}\right)^{1/q}, \qquad q \in Q\backslash\{0\},$$

$$F_0(s) = \exp\left\{\frac{1}{4N_s}\sum_{j=1}^{2N_s}\ln[F^2(j,s)]\right\}.$$

Шаг 5. Для каждого $q \in Q$ нужно выполнить линейную аппроксимацию зависимости $\ln\left[F_q(s)\right]$ от $\ln(s)$. При этом наклон полученной линейной функции – это оценка для $h(q)$. Масштабную функцию $\tau(q)$ получаем из выражения $\tau(q) = qh(q) - 1$.

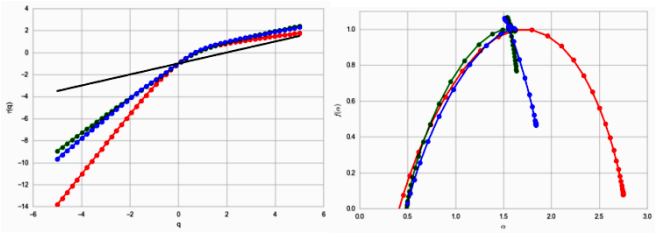
Шаг 6. Оценим производную полученной оценки функции $\tau(q)$

$$\alpha_0 = \left.\frac{d\tau(q)}{dq}\right|_{q=q_0}, \qquad q_0 \in Q,$$

и в результате получаем оценку мультифрактального спектра

$$f(\alpha) = q_0\alpha_0 - \tau(q_0).$$

Примеры масштабных функций и мультифрактальных спектров для рядов T, K и X полученные с помощью метода MF-DFA приведены на Рис. 43.



а б

Рис. 43 – Масштабные функции (а) и мультифрактальные спектры (б) для рядов T, K и X полученные с помощью метода MF-DFA

Для корректной работы приведенного алгоритма нужно заранее выбрать множество $\mathcal{S}$ длин отрезков, на которые делится ряд. Выясним, каким должно быть множество $\mathcal{S}$. С одной стороны, если использовать значения $s \geq N/4$, то мы разделим ряд на малое количество частей, и оценка $F_q(s)$ будет посчитана по малому количеству оценок дисперсий $F^2(j,s)$. С другой стороны, если использовать полиномиальную регрессию, то при $s \leq 10$ на шаге 3 будет использоваться мало точек для полиномиальной регрессии. Таким образом, разумное ограничение: $10 \leq s \leq N/4$. При этом в случае линейной регрессии, можно использовать и небольшие значения $s = 3, 4, \ldots$.

Для длинных временных рядов, которые содержат более двух тысяч значений, рекомендуется использовать ограничения

$$s_{min} = \max\{10, N/100\}, \qquad s_{max} = \min\{20s_{min}, N/10\},$$

и выбирать шаг таким образом, чтобы во множестве $\mathcal{S}$ было не более 100 значений. Также степень m нужно выбрать таким образом, чтобы регрессионный полином $X_j(i)$ адекватно отображал тренд в каждом куске временного ряда.

В случае коротких временных рядов, таких как T, K и X достаточно использовать линейную регрессию.

2.6.8. Сетевые модели

В последнее время выделилось отдельное научное направление – анализ социальных сетей (SNA, Social Networks

Analysis), которое базируется, с одной стороны, на социологии, а с другой на теории сложных сетей (Complex Networks) [Newman, 2003]. В рамках теории сложных сетей изучаются сетевые характеристики не только с точки зрения топологии сетей, но и статистические феномены, распределение весов отдельных узлов и ребер, эффекты протекания и проводимости. Несмотря на то, что в рассмотрение теории сложных сетей попадают различные сети (электрические, транспортные, информационные), наибольший вклад в развитие этой теории внесли исследования социальных сетей [Ландэ и др., 2009]. В теории сложных сетей выделяют три основных направления:

- исследование статистических свойств, которые характеризуют поведение сетей;
- создание моделей сетей;
- предсказание поведения сетей при изменении структурных свойств.

Параметры сетей

В прикладных исследованиях чаще всего применяются такие типичные для сетевого анализа характеристики, как размер сети, сетевая плотность, степень центральности и т. п. При анализе сложных сетей, как и в теории графов исследуются:

- параметры отдельных узлов;
- параметры сети в целом;
- сетевые подструктуры.

Для отдельных узлов выделяют следующие параметры:

- входная степень узла – количество ребер графа, которые входят в узел;
- выходная степень узла – количество ребер графа, которые выходят из узла;
- расстояние от данного узла до каждого из других;
- среднее расстояние от данного узла до других;
- эксцентричность (eccentricity) – наибольшее из геодезических расстояний (минимальных расстояний между узлами) от данного узла к другим;
- посредничество (betweenness), показывающее, сколько кратчайших путей проходит через данный узел;

- центральность – общее количество связей данного узла по отношению к другим.

Для анализа сети в целом используют такие параметры, как:

- число узлов;
- число ребер;
- геодезическое расстояние между узлами;
- среднее расстояние от одного узла к другим;
- плотность – отношение количества ребер в сети к возможному максимальному количеству ребер при данном количестве узлов;
- количество симметричных, транзитивных и циклических триад;
- диаметр сети – наибольшее геодезическое расстояние в сети и т.д.

Существует несколько актуальных задач математического исследования социальных сетей, среди которых можно выделить следующие основные:

Важной характеристикой сети является функция распределения степеней узлов $P(k)$, которая определяется как вероятность того, что узел i имеет степень $k_i = k$. Сети, характеризующиеся разными $P(k)$, демонстрируют различное поведение, $P(k)$ в некоторых случаях может быть распределениями Пуассона $\left(P(k) = e^{-m} m^k / k!\right)$, где m – математическое ожидание), экспоненциальным $\left(P(k) = e^{-k/m}\right)$ или степенным $\left(P(k) \sim 1/k^{\gamma},\ k \neq 0,\ \gamma > 0\right)$.

Сети со степенным распределением степеней узлов называются безмасштабными (scale-free). Именно безмасштабные распределения часто наблюдаются в реально социальных сетях. При степенном распределении возможно существование узлов с очень высокой степенью, что практически не наблюдается в сетях с пуассоновским распределением.

Расстояние между узлами определяется как количество шагов, которые необходимо сделать, чтобы по существующим ребрам добраться от одного узла до другого. Естественно, узлы могут быть соединены прямо или опосредованно. Крат-

чайшим путем d_{ij} между узлами i и j называется наименьшее расстояние между ними. Для всей сети можно ввести понятие среднего пути, как среднего по всем парам узлов кратчайшего расстояния между ними:

$$l = \frac{2}{n(n+1)} \sum_{i \geq j} d_{ij},$$

где n – количество узлов, d_{ij} – кратчайшее расстояние между узлами i и j.

Венгерскими математиками П.Эрдёшем (P. Erdős) и А. Реньи (A. Rényi) было показано, что среднее расстояние между двумя вершинами в случайном графе растет как логарифм от числа его узлов [Erdős, 1960].

Некоторые сети могут оказаться несвязными, т.е. найдутся узлы, расстояние между которыми окажется бесконечным. Соответственно, средний путь может оказаться также равным бесконечности. Для учета таких случаев вводится понятие среднего инверсного пути (его еще называют «эффективностью сети») между узлами, рассчитываемое по формуле:

$$il = \frac{2}{n(n-1)} \sum_{i > j} \frac{1}{d_{ij}}.$$

Сети также характеризуются таким параметром как диаметр или максимальный кратчайший путь, равный максимальному значению из всех d_{ij}.

Д.Уаттс (D.Watts) и С.Строгатц (S.Strogatz) в 1998 году определили такой параметр сетей, как коэффициент кластерности [Watts, 1998], который характеризует уровень связности узлов в сети, тенденцию к образованию групп взаимосвязанных узлов, так называемых клик (clique). Кроме того, для конкретного узла коэффициент кластеризации показывает, сколько ближайших соседей данного узла являются также ближайшими соседями друг для друга.

Коэффициент кластерности может определяться как для каждого узла, так и для всей сети. Для сети коэффициент кластерности определяется как нормированная по количеству

узлов сумма соответствующих коэффициентов для отдельных узлов.

Коэффициент кластерности для отдельного узла сети определяется следующим образом. Пусть из узла выходит k ребер, которые соединяют его с k другими узлами, ближайшими соседями. Если предположить, что все ближайшие соседи соединены непосредственно друг с другом, то количество ребер между ними составляло бы $1/2 \cdot k(k-1)$. Т.е. это число, которое соответствует максимально возможному количеству ребер, которыми могли бы соединяться ближайшие соседи выбранного узла. Отношение реального количества ребер, которые соединяют ближайших соседей данного узла к максимально возможному (такому, при котором все ближайшие соседи данного узла были бы соединены непосредственно друг с другом) называется коэффициентом кластерности узла $i - C(i)$. Естественно, эта величина не превышает единицы.

Посредничество (betweenness) – это параметр, показывающий, сколько кратчайших путей проходит через узел. Эта характеристика отражает роль данного узла в установлении связей в сети. Узлы с наибольшим посредничеством играют главную роль в установлении связей между другими узлами в сети. Посредничество b_m узла m определяется по формуле:

$$b_m = \sum_{i \neq j} \frac{B(i,m,j)}{B(i,j)},$$

где $B(i,j)$ - общее количество кратчайших путей между узлами i и j, $B(i,m,j)$ – количество кратчайших путей между узлами i и j, проходящих через узел m.

О «структуре сообщества» можно говорить тогда, когда существуют группы узлов, которые имеют высокую плотность ребер между собой, притом, что плотность ребер между отдельными группами – низкая. Традиционный метод для выявления структуры сообществ – кластерный анализ. Существуют десятки приемлемых для этого методов, которые базируются на разных мерах расстояний между узлами, взвешенных путевых индексах между узлами и т.п. В частности, для больших социальных сетей наличие структуры сообществ оказалось неотъемлемым свойством.

К свойствам реальных социальных сетей относятся и так называемые «слабые» связи. Аналогом слабых социальных связей являются, например, отношения с далекими знакомыми и коллегами. В некоторых случаях эти связи оказываются более эффективными, чем связи «сильные». Так, группой исследователей из Великобритании, США и Венгрии, был получен концептуальный вывод в области мобильной связи, заключающийся в том, что «слабые» социальные связи между индивидуумами оказываются наиболее важными для существования социальной сети [Bjorneborn, 2004].

Для исследования были проанализированы звонки 4.6 млн. абонентов мобильной связи, что составляет около 20% населения одной европейской страны. Это был первый случай в мировой практике, когда удалось получить и проанализировать такую большую выборку данных, относящихся к межличностной коммуникации.

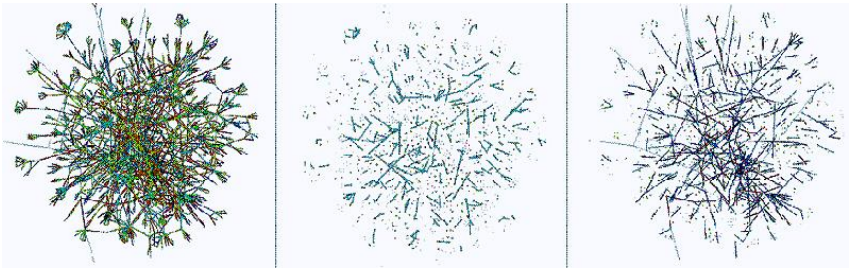
В социальной сети с 4.6 млн. узлов было выявлено 7 млн. социальных связей, т.е. взаимных звонков от одного абонента другому и обратно, если обратные звонки были сделаны на протяжении 18 недель. Частота и продолжительность разговоров использовались для того, чтобы определить силу каждой социальной связи.

Было выявлено, что именно слабые социальные связи (один-два обратных звонка на протяжении 18 недель) связывают воедино большую социальную сеть. Если эти связи проигнорировать, то сеть распадется на отдельные фрагменты. Если же не учитывать сильных связей, то связность сети нарушится (Рис. 44). Оказалось, что именно слабые связи являются тем феноменом, который связывает сеть в единое целое.

Несмотря на огромные размеры некоторых социальных сетей во многих из них существует сравнительно короткий путь между двумя любыми узлами – геодезическое расстояние. В 1967 г. психолог С. Милграм в результате проделанных масштабных экспериментов вычислил, что существует цепочка знакомств, в среднем длиной шесть звеньев, практически между двумя любыми гражданами США [Milgram, 1967].

Д.Уаттс и С.Строгатц обнаружили феномен, характерный для многих реальных сетей, названный эффектом малых ми-

ров (Small Worlds) [Watts, 1998]. При исследовании этого феномена ими была предложена процедура построения наглядной модели сети, которой присущ этот феномен.



1 2 3

Рис. 44 – Структура сети:

1) полная карта сети социальных коммуникаций;
2) социальная сеть, из которой изъяты слабые связи;
3) сеть, из которой изъяты сильные связи: структура сохраняет связность

Три состояния этой сети представлены на Рис. 45: регулярная сеть – каждый узел которой соединен с четырьмя соседними, та же сеть, у которой некоторые «ближние» связи случайным образом заменены «далекими» (именно в этом случае возникает феномен «малых миров») и случайная сеть, в которой количество подобных замен превысило некоторый порог.

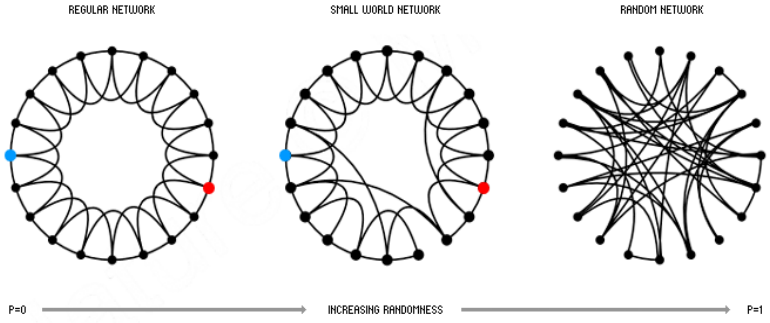


Рис. 45 – Модель Уаттса-Строгатца

В реальности оказалось, что именно те сети, узлы которых имеют одновременно некоторое количество локальных и случайных «далеких» связей, демонстрируют одновременно эффект малого мира и высокий уровень кластеризации.

На Рис. 46 приведены графики изменения средней длины пути и коэффициента кластеризации искусственной сети Д. Уоттса и С. Строгатца от вероятности установления «далеких связей» (в полулогарифмической шкале).

Например, WWW является сетью, для которой также подтвержден феномен малых миров. Анализ топологии веб, проведенный Ши Жоу (S.Zhou) и Р.Дж.Мондрагоном (R.J.Mondragon) из Лондонского университета, показал, что узлы с большой степенью исходящих гиперссылок имеют больше связей между собой, чем с узлами с малой степенью, тогда как последние имеют больше связей с узлами с большой степенью, чем между собой. Этот феномен был назван "клубом богатых" (rich-club phenomenon). Исследование показало, что 27% всех соединений имеют место между всего 5% наибольших узлов, 60% приходится на соединение других 95% узлов с 5% наибольших и только 13% - это соединение между узлами, которые не входят в лидирующие 5%.

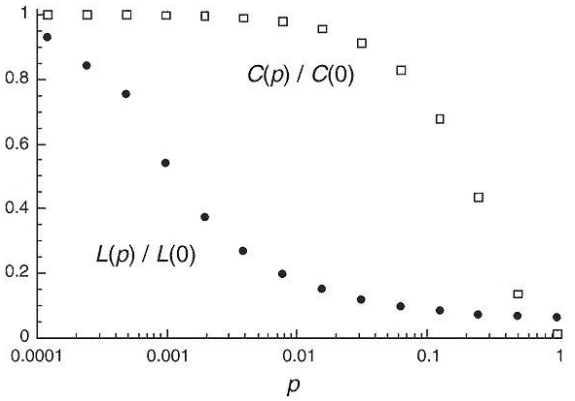


Рис. 46 – Динамика изменения длины пути и коэффициента кластерности в модели Уоттса-Строгатца в полулогарифмической шкале (ось 0X – вероятность замены ближних связей далекими)

Эти исследования дают основания полагать, что зависимость WWW от больших узлов значительно существеннее, чем предполагалось ранее, т.е. она еще более чувствительна к злонамеренным атакам. С концепцией «малых миров» связан также практический подход, называемый «сетевой мобилизацией», которая реализуется над структурой «малых миров». В частности, скорость распространения информации благодаря эффекту «малых миров» в реальных сетях возрастает на порядки по сравнению со случайными сетями, ведь большинство пар узлов реальных сетей соединены короткими путями.

Практикой доказано [Rothenberg, 2002], что террористические сети чаще всего не только безмасштабные, но также демонстрируют свойства малых миров, т.е. то, что наличие тесно связанных кластеров (групп тесно связанных узлов) обеспечивает локальная связь даже в случаях удачных атак, когда концентраторы (наибольшие посредники) выходят из строя.

При изучении «малых миров» определился интересный подход, логически связанный с понятием перколяции (протекания) [Broadbent, 1957], [Снарский, 2007]. Оказывается, что многие вопросы, которые возникают при анализе сетевой безопасности в Интернет, непосредственно относятся к этой теории. Самая простая, очищенная от всех физических и математических наслоений формулировка задачи теории перколяции имеет такой вид: «Данна сеть, случайная часть ребер которой проводит сигнал, а другая часть - не проводит. Основной вопрос - чему равна минимальная концентрация проводящих связей, при которой еще существует путь через всю сеть?». К задачам, которые решаются в рамках теории перколяции и анализа сетей относятся такие, как определение предельного уровня проводимости, изменения длины пути и его траектории при приближении к предельному уровню проводимости, количества узлов, которые необходимо вывести из строя, чтобы нарушить связнос сети.

Экспертами по проблемам безопасности эффект «малых миров» в последнее время все чаще связывается с сетями террористических организаций, так называемыми оверлейными сетями, т.е. сетями, надстроенными поверх сети Интернет.

Анализируя связи в сети, можно узнать о ее важных свойствах, например, выявить наличие кластеров, определить их состав, различия в связности внутри и между кластерами, идентифицировать ключевые элементы, связывающие кластеры между собой и т.д. Вместе с тем серьезным препятствием при анализе является неполная информация о связях между отдельными узлами сети.

Недавно группа исследователей из Института Санта Фе (Santa Fe Institute) представила алгоритм, с помощью которого становится возможным автоматическое получение информации об иерархической структуре подобных сетей [Clauset, 2008]. Новый метод восстановления связей может поступить на вооружение, как спецслужб, так и подразделений конкурентной разведки компаний. Так, например, зная только о половине связей между террористами, можно будет с высокой вероятностью восстановить недостающие звенья всей цепочки.

Даже не имея полного описания системы можно получать репрезентативную выборку связей и по ней пытаться достроить всю сеть. Анализ получающегося графа позволяет выявить потенциально важные связи, которые не удалось обнаружить в реальной сети. Например, имея информацию только о половине контактов участников сети между собой, можно с вероятностью 0,8 прогнозировать те связи, о которых сначала ничего не было известно. Очевидно, что данный метод может быть очень полезным при выявлении скрытых сетевых группировок, и таким образом поставить дело обеспечения государственной и коммерческой безопасности на качественно новый уровень.

Для анализа сложных сетей понятий, упоминаемых в отдельных документах из информационных потоков, могут применяться методы глубинного анализа текстов, а точнее контент-мониторинга и экстрагирования таких понятий, как персоны, компании, топонимы (географические названия) и т.п.

Одним из направлений анализа социальных сетей является визуализация, которая имеет важное значение, поскольку зачастую позволяет делать важные выводы относительно характера взаимодействия субъектов-узлов, не прибегая к точ-

ным методам анализа. При отображении модели сети может оказаться целесообразным:

- размещение узлов сети в двух измерениях;
- пространственное упорядочение объектов в одном измерении в соответствии с некоторыми количественными свойствами;
- использование общих для всех сетевых диаграмм методов для отображения количественных и качественных свойств объектов и отношений.

Сетевые признаки информационных операций

Как расширение приведенной выше мультиагентной модели распространения информации можно рассматривать модель, в которой учитывается структура формируемой сети [Пугачев, 2015]. В рамках этой модели каждый агент – источник информации – обладает не «потенциалом», а некоторым рейтингом (которому на схемах соответствует размер соответствующего узла). Связями в рассматриваемой сети являются факты перепечатки или «пересказа».

В основе модели лежит предположение, что при проведении информационных операций наиболее рейтинговые источники перепечатывают информацию у наименее рейтинговых, или образуются кластеры низкорейтинговых изданий, перепечатывающих одну и ту же новость.

На Рис. 47 приведен пример типичных информационных операций, выявляемых в рамках данной модели.

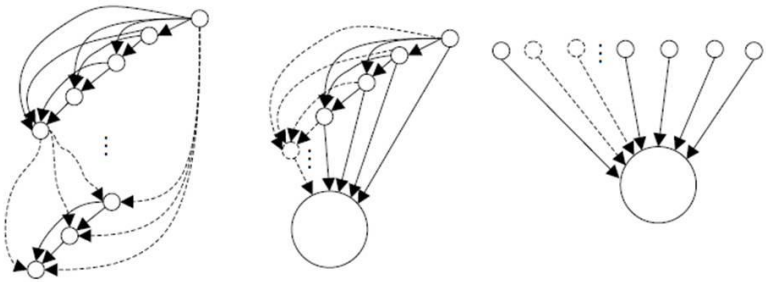


Рис. 47 – Примеры сетей распространения информации, имеющих признаки информационных операций

В рамках формализации этой же модели выбирается несколько десятков параметров топологии сетей распространения информации, таких как диаметр, плотность, кластеризация, посредничество и т.п., которые сравниваются с некоторыми эталонными значениями.

К достоинствам данной модели следует отнести ее формальную строгость и соответствие активно развивающемуся в последнее время направлению Complex Networks, что позволяет ожидать ее дальнейшего развития. К недостаткам следует, по-видимому, отнести малую корреляцию с содержательной стороной распознаваемых информационных операций, а также определенную вычислительную сложность при выявлении нечетких информационных дубликатов документов.

Технологические этапы исследования взаимного влияния источников информации

Для эффективного исследования взаимного влияния источников информации из сети Интернет (веб-ресурсов, социальных медиа) предлагается последовательность шагов, этапов обработки информации, каждый из которых сам по себе обеспечивает получение аналитического продукта. Совокупность таких этапов, которые базируются на использовании необходимых и доступных инструментальных средств, специальных приемов, можно рассматривать как процедуру проведения действий, направленных на получение аналитических материалов, включающих построение и анализ сети их взаимного влияния.

При проведении данных информационно-аналитических исследований на базе контент-мониторинга к таким задачам можно отнести:

- Нахождение релевантных публикаций на заданную тематику.
- Выявление взаимных контекстных ссылок и перепечаток в документах, представленных различными информационными источниками.
- Построение сети влияния, анализ и визуализация взаимосвязей информационных источников, в том числе ранжирование узлов построенной сети по степени влиятельности.

- Выявление возможных информационных операций и построение сценария противодействия информационным операциям в сетевой среде.

Получение репрезентативного массива публикаций

Для получения репрезентативного массива публикаций по выбранной тематике необходимо выбрать систему контент-мониторинга, которая предоставляет поток информационных сообщений по определенной тематике. Тематика может выражаться запросом на языке информационно-поисковой системы.

В качестве системы контент-мониторинга авторами была выбрана система InfoStream, которая в настоящее время охватывает 10 тыс. источников информации на русском и украинском языках. В базы данных системы ежедневно поступает более 100 тыс. документов. Система InfoStream обеспечивает поиск, а также просмотр списка и полных текстов релевантных документов.

В приведенном на Рис. 48 примере показан фрагмент интерфейса системы, через который отрабатывался запрос, относящийся к обсуждению в январе 2016 года вопрос отставки премьер-министра Украины А. Яценюка.

В результате был сформирован тематический информационный массив, который охватывает свыше 3 тыс. документов.

Определение контекстных ссылок

Основой построения сети влияния источников информации является контекстные ссылки и перепечатки в тематическом информационном потоке. Контекстные ссылки выявляются путем идентификации шаблонов в документах выбранного информационного массива и признаков точных перепечаток, определяемых методами выявления плагиата. В свою очередь, сами шаблоны периодически определяются/дополняются экспертами в автоматизированном режиме путем анализа контекста потока документов системы контент-мониторинга методами Text Mining.

Построение сети влияния источников информации

Найденные в текстах контекстные ссылки и перепечатки позволяют сформировать матрицу цитирования, транспонируя которую в соответствии с приведенной выше гипотезой, можно сформировать матрицу влияния. Данной матрицы соответствует сеть влияния источников, пример визуализации которой для рассмотренного выше тематического информационного массива с помощью системы Gephi приведен на Рис. 49.

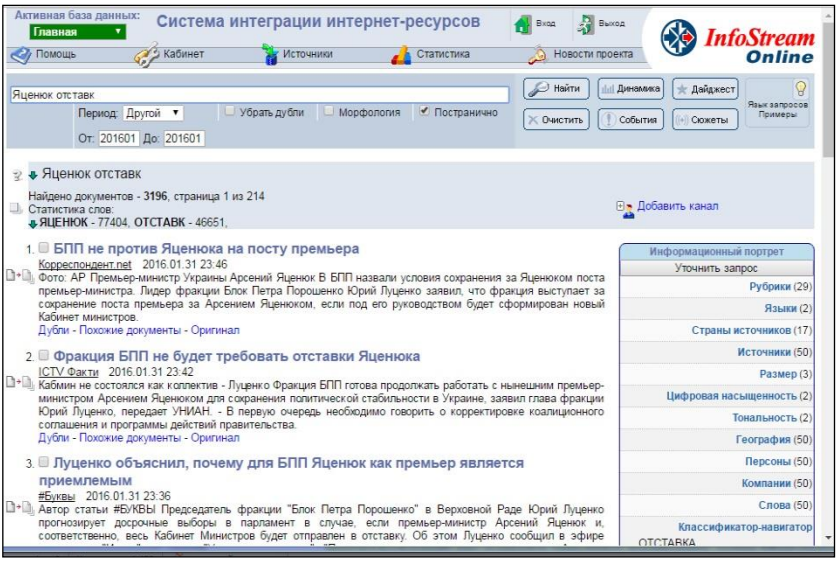


Рис. 48 – Фрагмент интерфейса системы контент-мониторинга

Исследование сети влияния источников информации

Построенную сеть влияния источников информации можно исследовать с помощью общепринятых инструментальных средств (например, с помощью системы Gephi были получены следующие параметры построенной сети, как количество узлов: 141, ребер 196, плотность графа: 0.01, средний коэффициент кластеризации: 0,026, средняя длина пути: 1,26 и т.д.).

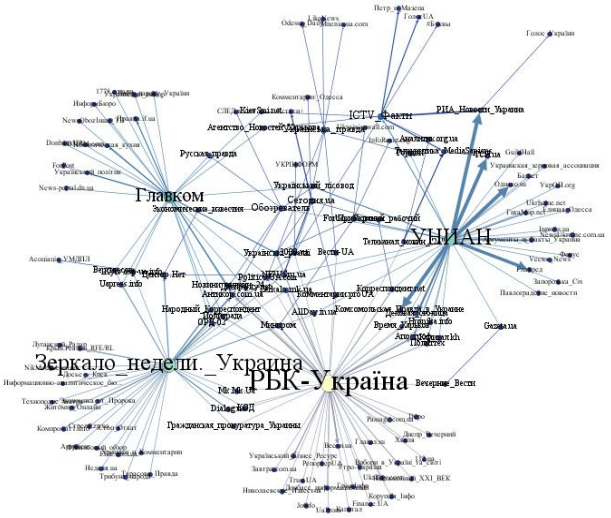


Рис. 49 – Фрагмент сети связей источников по выбранной тематике

Для содержательного анализа большое значение имеет вес узлов сети. Был получен такой список наиболее важных узлов по критерию выходной мощности:

№	Веб-ресурс	Выходная степень
1	РБК-Україна	50
2	Зеркало недели	38
3	УНИАН	35
4	Главком	28
5	ICTV-Факты	10
6	Сегодня.ua	7
7	Українська правда	7
8	Обозреватель	6
9	Forbes-Украина	4
10	Цензор.Нет	3

Перспективным подходом к ранжированию источников по уровню влияния является алгоритм HITS, предложенный Дж. Кляйнбергом [Kleinberg, 2006].

Алгоритм HITS обеспечивает выбор из сети лучших «авторов» (узлов, на которые введут ссылки) и «посредников» (узлов, от которых идут ссылки включения).

В соответствии с алгоритмом HITS для каждого узла сети рекурсивно вычисляется его значимость как автора $a(v_j)$ и посредника (хаба) $h(v_j)$ по формулам:

$$a(v_i) = \sum_{j \to i} h(v_j);$$

$$h(v_i) = \sum_{i \to j} a(v_j).$$

В этих формулах суммирование производится по всем узлам, которые ссылаются на (или на которые ссылается) данный узел.

Перефразируя обозначения, приведенные в [Kleinberg, 2006], а именно заменяя «авторство» на «подверженность воздействию», а «посредничество» на «влиятельность» можно с небольшими вычислительными затратами вычислять соответствующие характеристики узлов сети влияния.

Также для выявления информационных воздействий большое значение имеет определение «скрытых» связей. Методика определения скрытых связей, скрытых воздействий, в частности, приведена в работе [Snarskii, 2016].

Определение возможных информационных операций

Сеть информационного воздействия источников информации позволяет оперативно идентифицировать возможные информационные операции в соответствии с подходами, предложенными в [Потемкин, 2015]. Предполагается, что вероятность наличия информационной операции мала, если информация о происшествии сначала зарождается во влиятельном информационном источнике, а затем перепечатывается (со ссылками или без них) менее влиятельными источниками (Рис. 50). Обратные явления, когда более влиятельные

издания перепечатывают информацию в менее влиятельных, пусть и многочисленных, может быть признаком информационной операции, атаки (Рис. 51). Именно такие картины наблюдались при сетевом анализе реальных тематических информационных потоков (Рис. 52).

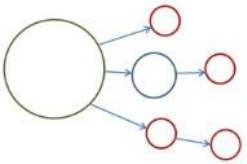


Рис. 50 – Типовой сценарий распространения информации

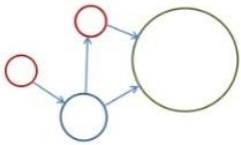


Рис. 51 – Сценарий распространения информации, присущий информационной операции

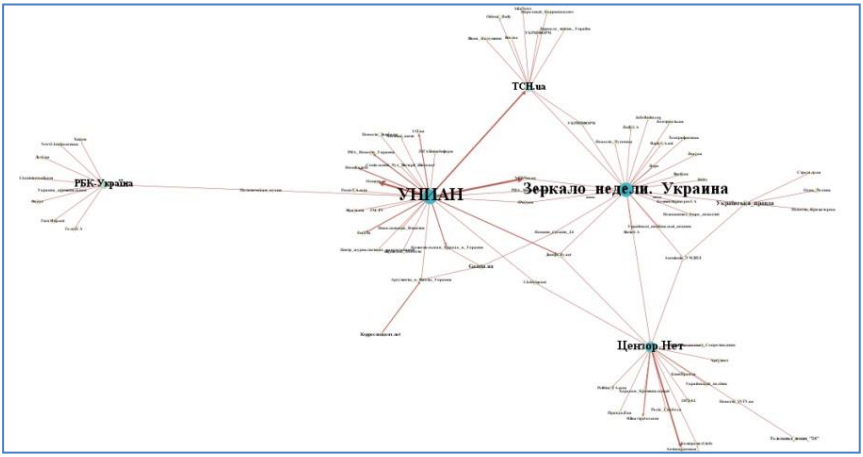


Рис. 52 - Фрагмент сети связей источников по специальной тематике

К преимуществам данной модели следует отнести ее формальную строгость и соответствие активно развивающемуся в последнее время направления Complex Networks, что позволяет ожидать дальнейшего развития. К недостаткам следует, по-видимому, отнести малую корреляцию с содержательной стороной распознаваемых информационных операций, а также определенную вычислительную сложность при обнаружении нечетких информационных дубликатов документов.

2.7. Реализованные технологии конкурентной разведки

Система конкурентной разведки должна позволять руководству, аналитическому, маркетинговому отделам компании не только оперативно реагировать на изменения ситуации на рынках, но и оценивать риски и возможности, прогнозировать их и принимать решения о дальнейших путях развития, обеспечить переход от традиционного интуитивного принятия решений на основе недостаточной информации к управлению, основанному на достоверных прогнозах и знаниях.

Одним из основных общих требований к системе конкурентной разведки должно быть соответствие цикла обработки информации в такой системе классическому информационному разведывательному циклу. Т.е. система должна самостоятельно или с участием оператора обеспечивать:

- выбор тематики и направлений разведки (целеуказание);
- выбор источников информации (веб-сайты, блоги, форумы и т.д.);
- автоматический поиск и скачивание информации по заданным направлениям мониторинга и указанным источникам по запланированному расписанию (планирование и сбор данных);
- обработку собранных данных и превращение их в информацию;
- контент анализ и синтез информации – превращение ее в знания;

- своевременную доставку информации к конечным потребителям.

Так как в целях конкурентной разведки необходимо анализировать данные из всех доступных источников информации, в которых эта информация может быть представлена в различных видах и форматах, то крайне важным требованием к системе является обеспечение ею единого информационного пространства взаимосвязанных объектов и фактов независимо от типа их источников или контента. Два других требования касаются сохранения связи объектов и фактов с релевантными данными и источниками информации (аргументированность) и обеспечения исторически-пространственной модели банка данных системы, что предполагает наличие у всех объектов атрибутов времени, места и источника данных, а также невозможность их безвозвратного удаления из системы с течением времени.

Основными объектами учета и мониторинга в системах конкурентной разведки, как правило, являются:

- источники информации (официальные сайты, интернет-издания, персональные сайты организаций или лиц, Интернет представительства печатных СМИ, информагентств, теле- и радиоканалов, открытые базы данных, объекты учета и т.д.);
- географические регионы;
- рынки и направления бизнеса;
- структуры (предприятия, организации и т.д.);
- персоны (конкуренты, контрагенты, партнеры, сотрудники, кандидаты и т.д.);
- нормативно-законодательная база и факты ее нарушения;
- политико-экономическая ситуация;
- криминальная обстановка;
- другие специализированные тематики.

Безусловно, система конкурентной разведки, использующая Интернет как один из источников информации, должна настраиваться под специфику деятельности компании. Она должна включать в себя соответствующую классификацию, гибкие механизмы поиска, оперативной доставки данных, а также качественной оценки информации. Одной из самых

важных задач анализа информации является определение ее достоверности, т.е. задача анализа и фильтрации шума и ложной информации. Без таких оценок всегда есть риск принять неверные решения. После анализа достоверности информации должны следовать оценки ее точности и важности. Главным критерием достоверности данных на практике является подтверждение информации другими источниками, заслуживающими доверия.

Даже поверхностный анализ основных требований к системам конкурентной разведки в Сети, показывает, что традиционные поисковики в системе Интернет не могут считаться полноценными инструментами конкурентной разведки в Интернет.

Информационные системы конкурентной разведки можно также условно классифицировать по наличию в них модулей автоматического и экспертного извлечения фактов. Соотношение между автоматически извлекаемыми системой и вручную (с помощью экспертов) фактами, событиями, объектами учета в разных системах разное. Автоматически извлекаемые системой факты называют А-фактами, факты, извлекаемые экспертами, – Э-фактами [Киселев, 2005].

Существующие на рынке системы конкурентной разведки отличаются как по своей полноте и соответствию полному разведциклу, так и своему инструментарию и соответственно своей цене. Кроме того, системы могут быть предназначены для использования в качестве инструментария исключительно собственными силами внутреннего подразделения конкурентной разведки предприятия, либо предполагать вынесение части задач на аутсорсинг специализированными структурами конкурентной разведки. Выбор систем, подходов и методик конкурентной разведки остается за потребителем, и в каждом случае индивидуален. Да это и понятно, нельзя же сравнивать потребности и выполняемые задачи аналитика спецслужбы и сотрудника, к примеру, маркетингового отдела малого предприятия.

В настоящее время в мире существует ряд систем, которые частично реализуют решения приведенных выше задач мониторинга субъектов, извлечения фактов, построения связей, однако некоторые из них не выдерживают критики по функциональности, некоторые имеют слишком высокую це-

ну. Кратко остановимся на возможностях некоторых подобных систем, реализованных в настоящее время.

Система ***RCO*** (www.rco.ru) – основное назначение – выявление фактографической информации из неструктурированных текстов больших объемов (поиск фактов в Big Data). Обладает широким спектром алгоритмов и технологий интеллектуальной обработки текстов, представленных на естественном языке. В частности, технологии RCO позволяют решать задачи выявления именованных объектов, связей и фактов из массивов неструктурированных данных. RCO Fact Extractor – это персональное приложение для Windows, которое предназначено для аналитической обработки текста на русском языке и выявления фактов, связанных с заданными объектами – лицами и организациями. Основная сфера применения программы – это задачи из областей конкурентной разведки, борьбы с коррупцией, требующие высокоточного поиска информации.

RCO Zoom (Рис. 62) – специализированная поисково-аналитическая система, сочетающая функционал традиционных поисковых систем с функциями контент-анализа в реальном времени и транзакционного хранилища документов. RCO Zoom обладает инструментарием для проведения эффективного оперативного поиска и аналитических исследований информации.

Система RCO Zoom позволяет работать с огромными массивами текстовой информации в реальном времени (объем базы – до сотен гигабайт, время поиска и обработки – секунды). Средство отображения – информационный портрет дает возможность получить ключевые слова, формулировать и проверять гипотезы, разделять объекты, выделять статистические инварианты в первом приближении. Систему можно использовать в качестве высоконадежного хранилища документов. Система может работать с документами на разных языках. Она интегрирована с библиотекой RCO Fact Extractor SDK. Интерфейс для языка Python позволяет реализовывать всевозможные надстройки для решения совершенно различных задач помимо хранения и поиска: от нахождения информационных дублей документов до их классификации и кластеризации.

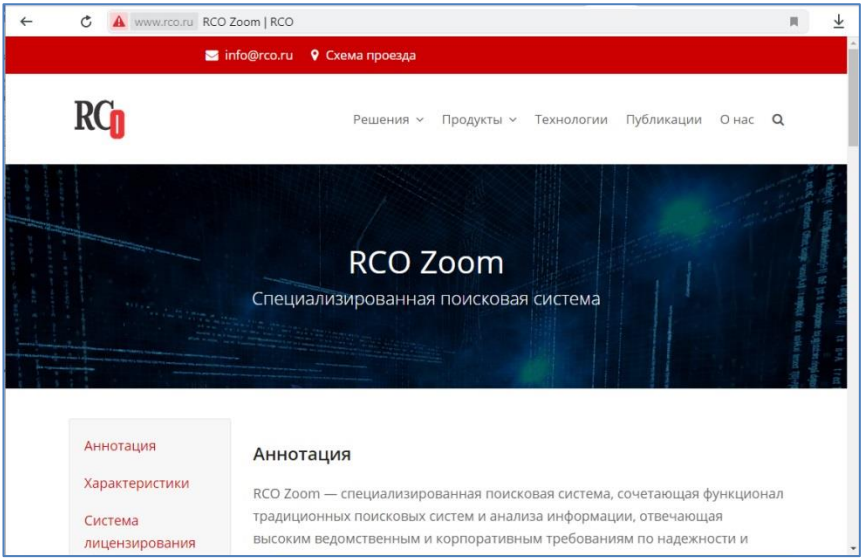


Рис. 62 – Фрагмент веб-сайта с описанием возможностей системы RCO Zoom

«Медиалогия» (www.mlg.ru) – сервис, обеспечивающий онлайновый доступ к базе СМИ с возможностью производить автоматический мониторинг СМИ и экспресс-анализ полученных сообщений в режиме реального времени. Система, на которой строится сервис, состоит из двух основных частей: база данных: 56 000 источников СМИ, 900 млн аккаунтов социальных сетей (по данным на 01.06.2020); автоматизированный аналитический модуль. База СМИ «Медиалогии» обновляется и пополняется на ежедневной основе. С помощью «Медиалогии» можно осуществлять оперативный мониторинг СМИ компании, ее топ-менеджеров, брендов, конкурентов и др. Возможности фильтров позволяют настроить мониторинг и оценивать тональность прессы, журналов, ТВ и интернет-изданий практически под любые информационные задачи.

PolyAnalyst (www.megaputer.ru) – основной продукт компании Megaputer Intelligence. PolyAnalyst - это программная платформа для анализа данных предоставляет среду для интеллектуального анализа текста, интеллектуального анализа данных, машинного обучения и прогнозной аналитики. Гра-

фический пользовательский интерфейс PolyAnalyst содержит различные узлы, которые можно связать в блок-схему для выполнения анализа. Программное обеспечение предоставляет узлы для импорта данных, подготовки данных, визуализации данных, анализа данных и экспорта данных. Функции текстовой аналитики PolyAnalyst включают узлы для кластеризации текста, анализа тональности, извлечения фактов, ключевых слов и сущностей, а также создания таксономий и онтологий. Polyanalyst также содержит узлы для анализа структурированных данных и выполнения кода на Python и R. По состоянию на 2020 год программное обеспечение поддерживает анализ текста на 16 языках. PolyAnalyst обычно используется для создания специализированных инструментов для бизнеса. Он использует модель клиент-сервер и лицензируется по модели «программное обеспечение как услуга».

IBM Integrated Analytics System (https://www.ibm.com/ru-ru/products/integrated-analytics-system) – комплексное решение для анализа гибридных данных, обеспечивающее массовую параллельную обработку. Решение включает в себя аппаратную платформу и программную систему запросов к базе данных для поддержки анализа различных типов данных. Предоставляемые готовые решения настроены и протестированы. В системе обеспечивается доступ к данным, запрос и анализ данных в хранилище данных и Hadoop режиме реального времени с помощью средств машинного обучения, обеспечивается возможность разрабатывать и совершенствовать модели машинного обучения непосредственно на платформе хранения данных.

Modus BI (https://modusbi.ru/) – Платформа для бизнес-аналитики, позволяющая собирать и визуализировать данные из различных источников, формировать отчетность и создавать прогнозы для принятия эффективных управленческих решений и мониторинга наиболее важной информации для бизнеса. Система Modus собирает данные из разнородных источников, очищает и готовит их для анализа со скоростью 50 млн. строк в час. Решение Modus ETL и Data Quality Management. Позволяет собирать данные из множества источников, обеспечивает процессы верификации, нормализации и последующего формирования единого корпоративного хранилища данных.

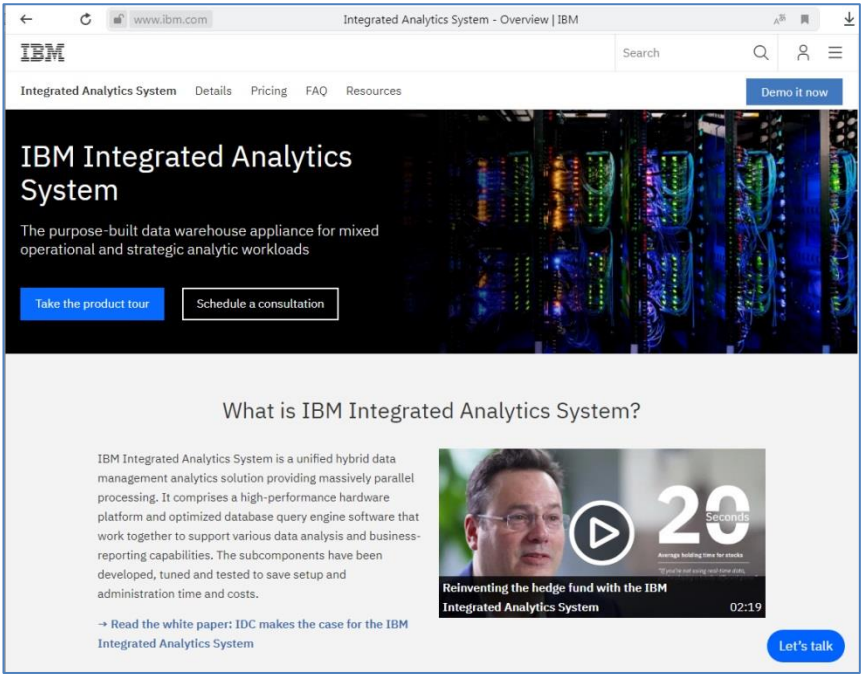


Рис. 63 – Фрагмент веб-ресурса IBM Integrated Analytics System

Oracle Analytics Cloud (https://www.oracle.com/business-analytics/analytics-cloud.html) – интегрированный комплекс аналитических инструментов, который разработан с целью обеспечить лучшее видение и понимание бизнеса широкому кругу пользователей и позволяет любому пользователю организации получить быстрый веб-доступ к актуальной информации. В системе Oracle Analytics Cloud встроенные средства машинного обучения могут дополнять исходные данные и предлагать варианты их интеллектуального обогащения. Средства машинного обучения включают элементы автоматического объяснения, что обеспечивает упрощение анализа данных и создание прогнозной аналитики. Встроенное машинное обучение позволяет формировать прогнозы, основанные на данных.

SAP BusinessObjects (https://www.sap.com/index.html#business-process-intelligence) – Гибкая, масштабируемая сис-

тема бизнес-аналитики (BI), которая позволяет находить и обмениваться данными для эффективного принятия решений. SAP BusinessObjects Business Intelligence - это централизованный пакет для создания отчетов, визуализации и совместного использования данных в режиме реального времени. Система преобразует исходные данные в полезную и доступную аналитическую информацию, доступную в любое время и в любом месте.

Rocket Folio/NXT (https://www.rocketsoftware.com/products/rocket-folionxt) – программы, позволяющие выявлять элементы информации, такие как сущности, взаимные связи и события, в неструктурированных текстах, а также выявлять неявные взаимосвязи и события в текстах. Rocket Folio, платформа на основе приложений, обеспечивает надежный поиск, управление контентом и его публикацию. Rocket NXT, серверная платформа поиска и публикации, предоставляет интегрированное решение как для структурированных, так и для неструктурированных данных — с любого устройства, в любом месте, в любое время.

В последнее время все основные западные бренды, специализирующиеся на разработке хранилищ и баз данных, корпоративных системах управления, расширили свои линейки продуктов модулями Business Intelligence (BI) или, дословном переводе – деловой разведки. О наличии таких модулей заявляют Oracle, SAS, SAP, IBM и другие бренды.

По заказу группы аналитиков Гарвардского университета российские разработчики из компании «Инфорус» создали информационно-аналитическую систему (ИАС) ***Avalanche*** (www.tora-centre.ru/avl3.htm), предназначенную для мониторинга изменений, происходящих в Интернете. Она собирает информацию с веб-страниц по заданному алгоритму и складывает эту информацию в собственную базу данных.

Технология Avalanche базируется на трех компонентах: автономном интеллектуальном поисковом роботе, создании «умных» папок и встроенной базе данных, позволяющей преобразовать их в «персональную энциклопедию». При работе с ИАС Avalanche формируется модель, требуемой пользователю области в виде набора «умных папок», каждая из которых «знает», что должно в нее попадать, и обеспечивает отсутствие дублирования. Наполнением «умных» папок занимается

специализированный поисковый робот, который запускается с компьютера в соответствии с установленными, требуемыми пользователю настройками. Робот может запускаться и автоматически в определенное установленное для него время. В Avalanche предусмотрены «тонкие» настройки, которые позволяют производить более детальный мониторинг.

«***Семантический архив»*** (www.anbr.ru) – аналитический инструмент, позволяющий автоматизировать всю технологическую цепочку решения аналитических и разведывательных задач, начиная от сбора необходимой информации, ее интеллектуального анализа и заканчивая удобным представлением отчетов. Платформа дает возможность анализировать и применять разнородную информацию для своевременного принятия оптимальных управленческих и бизнес-решений. «Семантический архив» имеет модульную структуру, что позволяет легко подобрать и настроить нужную конфигурацию системы.

Гибко настраиваемая онтологическая модель данных позволяет работать с разными тематиками и сферами деятельности. ИАС «Семантический архив» позволяет хранить информацию, импортированную из различных реляционных баз данных, вводить информацию из любых других источников: Интернет, СМИ, базы данных, онлайн библиотеки и системы (Спарк, Интегрум и др.), любой документ, собственные сведения экспертов.

Созданное хранилище служит аналитикам для поиска информации, добавления конфиденциальных собственных данных, выявления взаимосвязи между объектами и событиями, получения аналитических отчетов, визуализации: схем, графиков и карт.

Система управления досье ***X–Files – продукт компании*** «Ай-Теко» (https://www.i-teco.ru/), предназначенный для решения задачи выделения достоверных фактов из различных источников, заполнения ими досье на объекты мониторинга и их последующей аналитической обработки. X-Files — система управления досье, предназначенная для анализа фактов из различных источников. Система помогает принимать решения при наличии большого объема информационных «пробелов» в формировании целостного образа объекта. Она используется для обеспечения процессов принятия решений при

наличии большого объема «сырого» контента, что характерно для деятельности органов государственной власти, правоохранительных органов, крупных коммерческих компаний.

Система X–Files предполагает реализацию трех принципов:

1) единое информационное пространство взаимосвязанных фактов или гипотез независимо от типа их контента (содержимого источников информации);

2) связь фактов или гипотез с релевантными источниками информации (аргументированность фактов и гипотез);

3) исторически-пространственная информационная модель базы данных фактов и гипотез. Это означает наличие атрибутов времени и места для каждого факта, а также невозможность их безвозвратного удаления из системы.

В Xfiles реализована семантическая сеть, отражающая лишь взаимосвязи между объектами.

IBM Security i2 Analyst's Notebook (https://www.ibm.com/products/i2-analysts-notebook) – система визуального проектирования структуры данных для хранения данных о различных персонах и организациях.

В базе данных предусматривается возможность хранения определенных событий, происходящих с ними и имеющиеся взаимосвязи. Система IBM Security i2 Analyst's Notebook позволяет быстро и эффективно проводить анализ системы взаимосвязанных объектов и динамики последовательных событий, отображая результаты исследования в виде удобных для понимания схем и диаграмм. Решение предоставляет такие функции, как визуализация подключенных сетей, анализ социальных сетей и геопространственные или временные представления, которые помогут вам обнаружить скрытые связи и закономерности в данных. Информация отображается на диаграмме в виде объектов, к которым при необходимости можно добавить дополнительные атрибуты и карточки данных с комментариями. Объекты на диаграмме могут представляться не только в виде пиктограмм, но и в виде фотографий, файлов, аудиозаписей, видеозаписей и т.д. Программа позволяет создавать диаграммы с помощью запросов к реляционным базам данных, а также импорта данных из внешних файлов. При помощи имеющихся в Analyst's Notebook функций можно объединять элементы диаграмм,

искать существующие между ними связи, использовать систему поиска элементов, прослеживать «путь», объединяющий объекты, и т.п.

Система Analyst's Notebook предоставляет целый ряд удобных форматов визуализации, каждый из которых по-своему проясняет смысл информации и демонстрирует связи между объектами. Analyst's Notebook снабжен редактором, позволяющим в графической форме сформулировать запрос для поиска объектов и выявления их связей, создавать шаблоны интересующих событий. Систему Analyst's Notebook можно интегрировать в уже работающие у пользователя приложения. Система обеспечивает:

- поиск общих элементов и взаимосвязей, скрытых в данных;
- простоту интерпретации сложной информации;
- графическое отображение результатов;
- создание динамичных диаграмм;
- распространение диаграмм в печатном и электронном виде.

Говоря о продуктах, лидирующих в области Business Intelligence, следует отметить, что под этим термином, как правило, понимается набор инструментальных средств анализа статистических цифровых данных и других корпоративных отчетов и их визуализации, в отличие от Competitive Intelligence (конкурентной разведки), которая является гораздо более широким направлением информационно-аналитической деятельности.

На украинском рынке в сегменте информационно-аналитических систем конкурентной разведки представлены такие системы, как X-SCIF, «Энциклопедия деловой информации Украины», «Iceberg BI», «Страбис-ВЭБ» и др.

Хотелось бы отметить, что далеко не все из названных систем имеют полный функционал и соответствующие модули, обеспечивающие выполнение всего спектра задач конкурентной разведки.

В качестве одной из наиболее полнофункциональных отечественных систем, обработка информации в которой соответствует классическому информационному разведывательному циклу, можно назвать систему X-SCIF.

Рассмотрим, как реализуются этапы разведцикла с помощью данной системы, для чего остановимся на описании возможностей системы X-SCIF чуть подробней.

Онлайновая инструментальная корпоративная система мониторинга, агрегации и анализа информации X-SCIF (далее – ИКС X-SCIF*)* представляет собой программно-технический комплекс, предназначенный для решения задач автоматизированного сбора, обработки, создания интегрированного банка данных и анализа разнообразной информации.

Система X-SCIF обеспечивает:

- мониторинг информации с заданных пользователем веб-сай-тов (веб-страниц) в сети Интернет (Интранет) по заданным темам;
- поиск новых источников информации в сети Интернет по заданным пользователем тематикам и их последующую постановку на мониторинг;
- создание и сохранение сложных запросов по заданным темам, в виде каталогизированного списка или рубрики, для последующего проведения автоматического мониторинга, поиска или контент-анализа;
- приведение отобранной информации к единому формату и ее загрузку в хранилище;
- фильтрацию, классификацию, кластеризацию, рубрицирование и анонсирование загруженной полнотекстовой информации;
- автоматическое экстрагирование (извлечение) из полученной информации сущностей (объектов и фактов);
- создание, на базе загруженной в систему неформализованной полнотекстовой и формализованной фактографической информации, интегрированного банка данных (хранилища) объектов, фактов, событий и документов, связанных между собой различными видами и мотивами связей, с учетом атрибутов достоверности, актуальности, а также, весовых коэффициентов таких связей;
- сквозной поиск информации по запросам или темам пользователя, охватывающий, как поиск по внутреннему интегрированному банку данных, ранее уже извлеченной и накопленной информации,

так и онлайновый метапоиск в сети Интернет (поисковые системы, веб-сайты, блоги, социальные сети) и другим подключенным внешним источникам данных (официальные банки данных госорганов, БКИ и т.д.);

- аналитическую обработку информации (позволяет анализировать совместное упоминание и выявлять неявные связи между обьектами, отождествлять объекты и группировать информацию по сюжетам, строить цепочки и графы связей, анализировать информационную активность, эмоциональную окраску документов, пересечение заданных рубрик или тем, автоматически создавать информационный портрет отобранных по запросу документов, выделяя упоминаемые в них обьекты, источники, регионы и т.д., вычислять индекс информационного благоприятствования и многое другое);
- генерацию выходных форм по заданным пользователем параметрам (позволяет автоматически создавать электронное досье, схемы связей, дайджесты, обзоры, сравнительные диаграммы, информационные справки и агрегированные отчеты);
- оперативную доставку результатов запросов по различным каналам (в состав системы входит виртуальный офис, с собственным удаленным криптозащищенным хранилищем документов и почтовой системой, что позволяет обеспечивать, как «онлайновый» безопасный доступ по шифрованному каналу к документам, хранящимся в облаке, так и «оффлайновое» получение результирующих документов по электронной почте).

Структурно ИКС X-SCIF состоит из нескольких подсистем, ориентированных на соответствующие потребности корпоративных заказчиков, а именно:

- X-Stream – подсистема мониторинга web-сайтов, созданная на основе технологии InfoStream, а также полнотекстового банка данных (архива) неформализованной информации (статей, сообщений и т.д.), который автоматически пополняется с 1996 года и

является наиболее полным среди существующих электронных архивов в Украине;

- X-Files – интегрированный банк данных для накопления разнообразной формализованной справочно-фактографи-ческой информации, экстрагированной и агрегированной из всех доступных системе источников информации, стоящих на мониторинге, а также системы сквозного поиска по внутренним и внешним источникам (веб-сайтам, блогам, онлайновым базам данных, социальным сетям и т.д.).
- X-Office – система виртуального офиса, обеспечивающая безопасный доступ к корпоративным ресурсам из любой точки мира без установки дополнительного программного обеспечения. Система включает в себя «облачное» файловое хранилище документов и защищенную корпоративную веб-почту. Дополнительно в виртуальный офис может быть интегрирован сервер VoIP-телефонии для ведения конфиденциальных переговоров.
- X-Scoring – предскоринговая система, которая позволяет в автоматическом режиме проводить верификацию данных и предварительную проверку благонадежности контрагентов (физических и юридических лиц).

Остановимся на рассмотрении каждой из подсистем более детально.

Подсистема X-Stream, построенная на основе технологии InfoStream компании ElVisti, предназначена для мониторинга информации в сети Интернет по заданным пользователем параметрам, поиска информации по запросам или темам пользователя, оперативной доставки результатов поиска, и, таким образом, минимизации усилий и экономии времени, потраченного на поиск и обработку необходимой информации. Подсистема X-Stream предоставляет пользователю доступ к информации по интересующей его тематике одновременно с большого числа веб-сайтов, включая и те избранные, которые он привык просматривать ежедневно. В настоящее время осуществляется автоматический мониторинг более 7000 источников, поток информации превышает 80 000 документов в сутки. Территориальный охват – русско- англо- и

украино-язычные издания Украины, России и других стран ближнего и дальнего зарубежья. При необходимости может быть охвачен любой веб-сайт, доступный в сети Интернет. Информация из системы никогда не удаляется, а переносится в архив. Архив публикаций ведется непрерывно с 1996 года и составляет в настоящее время свыше 85 миллионов документов.

Отличие данной подсистемы от конкурирующих продуктов – ее объемы и возможность индивидуальной настройки. Она ориентирована не только на быструю доставку общих лент новостей, которых много в веб-пространстве, но и на осуществление мониторинга по индивидуально заданными пользователем параметрам или архивного поиска.

Просмотр информации осуществляется через единый унифицированный интерфейс. Пользователь может в режиме реального времени не только получать результаты поиска, но и формировать дайджесты, информационные досье, строить сюжетные цепочки, анализировать взаимосвязь рубрик, информационную активность, информационные связи и совместное упоминание объектов и т.д.

Ниже приведены примеры вывода результатов поиска (рис. 64), просмотра отдельного материала (рис. 65) и результата аналитической обработки результатов поиска – дайджеста (рис. 66).

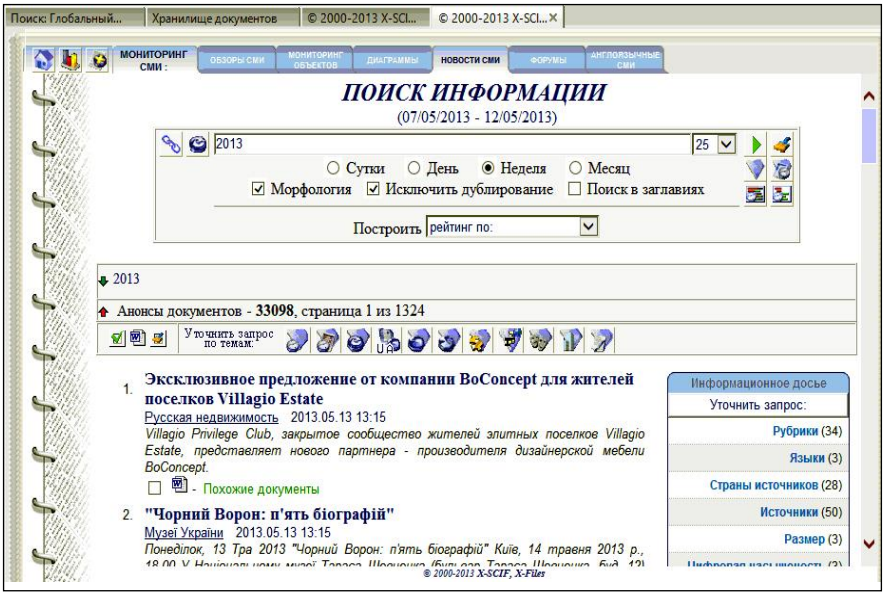


Рис. 64 – Вывод результатов поиска

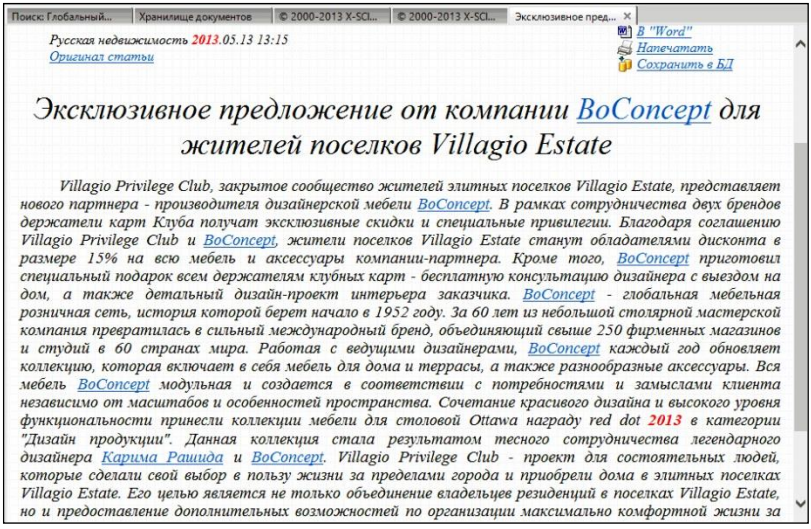


Рис. 65 – Просмотр статьи

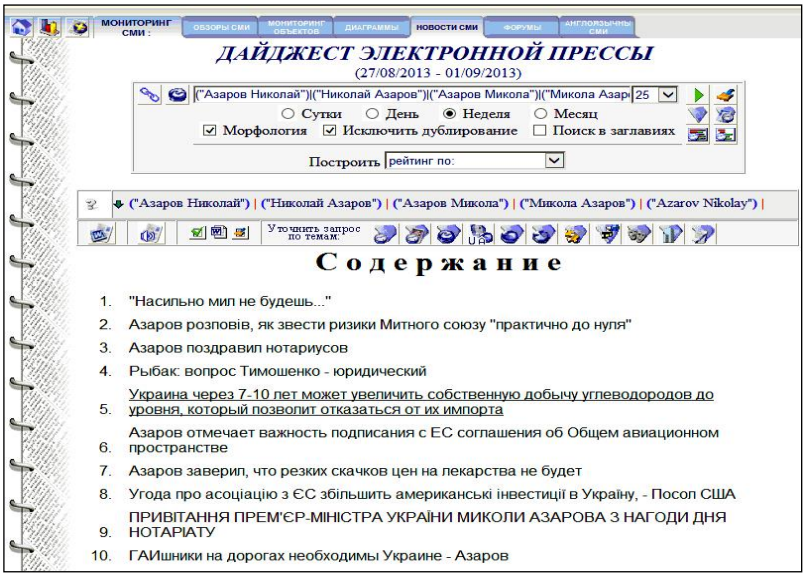


Рис. 66 – Дайджест

Использование подсистемы X-Stream позволяет:

- оперативно получать необходимую информацию по мере ее появления в Интернет, анализировать события, своевременно на них реагировать;
- формировать собственные информационные каналы, которые обусловлены запросами на информационно-поисковом языке, формировать архивы для последующей обработки и ретроспективного анализа;
- анализировать поток информации, поступающей в режиме реального времени;
- своевременно выявлять тенденции развития и состояние рынков товаров или услуг;
- отслеживать информацию о деятельности отдельных организаций, партий, движений, их PR-активность;
- оценивать возможные сферы влияния конфликтных или кризисных ситуаций, осуществлять информационный контроль вероятных источников рисков;

- находить и проверять потенциальных партнеров и клиентов.

Следующим структурным элементом ИКС X-SCIF является подсистема X-Files (не следует путать с известной российской системой). Данная подсистема предназначена для накопления и хранения формализированной информации, полученной из всех доступных источников, осуществления сквозного поиска и дальнейшей аналитической обработки найденной информации.

Информация, полученная из различных источников, обрабатывается, формализируется, приводится к единому виду и записывается в интегрированный банк данных, который структурно охватывает объекты и связи между ними. Его структура, разработанная с учетом практических потребностей аналитиков, включает в себя более 40 типов объектов учета и более 1000 мотивов связей между ними.

Поиск необходимой информации осуществляется средствами глобального поиска, который выполняется по всем доступным банкам данных, а также предусматривает автоматическое получение информации от онлайн-поставщиков информации.

На рис. 67 приведен интерфейс ввода запросов для глобального поиска. Подсистема позволяет представлять результаты поиска в различной форме, наиболее удобной для решения текущей задачи.

Одной из наиболее распространенных форм представления отобранных данных является информационное досье (рис. 68). Данная форма позволяет отображать информацию об объекте учета интегрированного банка данных в виде, в котором представлены все реквизиты данного объекта учета, а также все связанные с ним записи в других банках данных. Формат вывода информационного досье приведен на рис. 69.

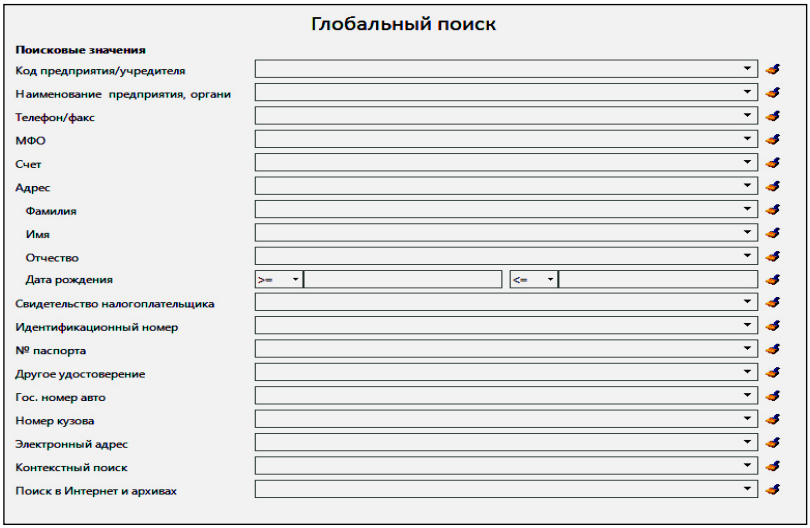


Рис. 67 – Интерфейс ввода запросов глобального поиска

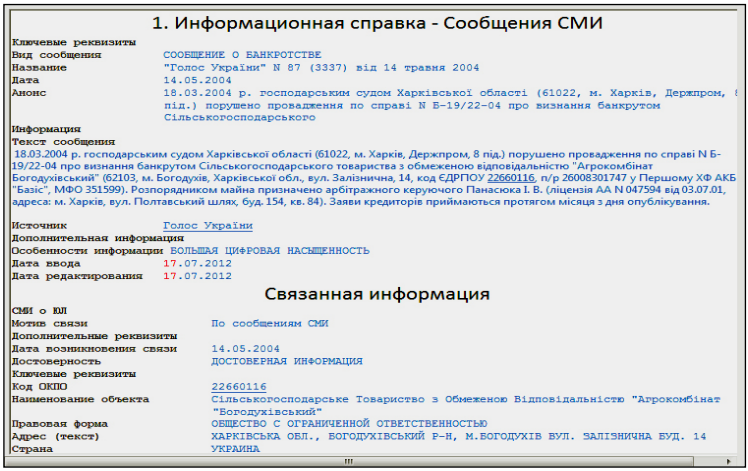
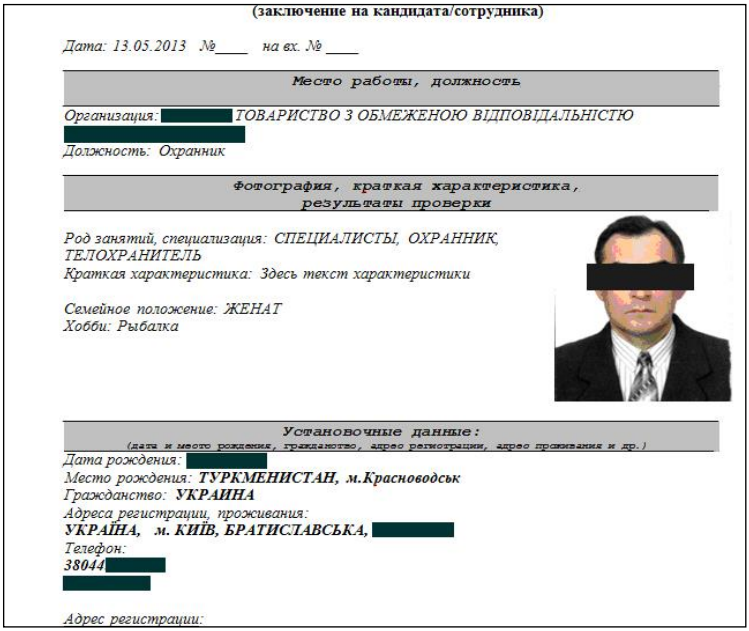


Рис. 68 – Информационное досье



(заключение на кандидата/сотрудника)

Дата: 13.05.2013 №____ на вх. № ____

Место работы, должность

Организация: [redacted] ТОВАРИСТВО З ОБМЕЖЕНОЮ ВІДПОВІДАЛЬНІСТЮ [redacted]

Должность: Охранник

Фотография, краткая характеристика, результаты проверки

Род занятий, специализация: СПЕЦИАЛИСТЫ, ОХРАННИК, ТЕЛОХРАНИТЕЛЬ

Краткая характеристика: Здесь текст характеристики

Семейное положение: ЖЕНАТ

Хобби: Рыбалка

Установочные данные:

(дата и место рождения, гражданство, адрес регистрации, адрес проживания и др.)

Дата рождения: [redacted]

Место рождения: **ТУРКМЕНИСТАН, м.Красноводськ**

Гражданство: **УКРАИНА**

Адреса регистрации, проживания:

УКРАЇНА, м. КИЇВ, БРАТИСЛАВСЬКА, [redacted]

Телефон:

38044 [redacted]

Адрес регистрации:

Рис. 69 – Формат вывода информационного досье

Еще одним примером формы представления данных является графическое досье. Отобранные объекты, вместе со своими связями, отображаются в виде графа, в котором вершинами выступают объекты учета, а ребрами – связи между соответствующими объектами (рис. 70). Такая форма представления позволяет осуществлять аналитические исследования как явных, так и неявных связей объектов учета, представлять их на экране в виде графов в различных масштабах, печатать схемы этих графов т.д. Также в данном режиме пользователю доступны все инструменты по редактированию, вводу и удалению информации, обеспечивающие интуитивное и быстрое редактирование формализированных данных.

Для аналитической обработки больших объемов однотипной информации в системе предусмотрен механизм агрегированных форм. Он позволяет на основании исходной информации, которая плохо поддается непосредственному ана-

лизу, строить агрегированные формы, графики, проводить расчет интегральных характеристик.

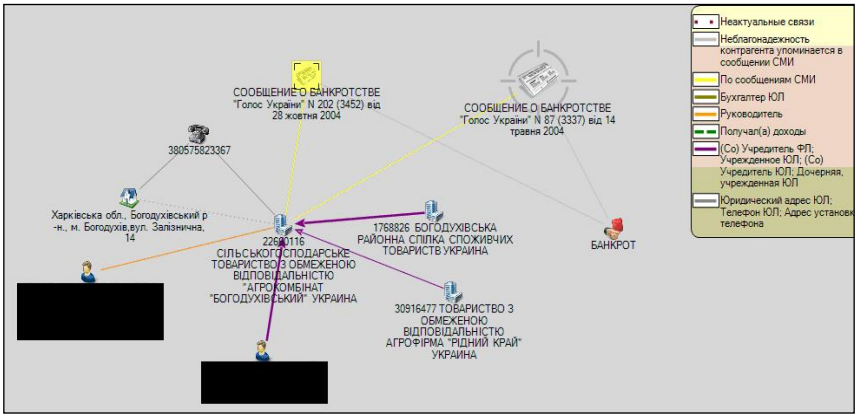


Рис. 70 – Визуализация графов связей объекта учета

Одной из ключевых возможностей подсистемы X-Files является модуль автоматизированного ввода и распознавания полнотекстовых документов. Он позволяет без участия оператора создавать объекты учета (лица, компании, телефоны, адреса, электронные адреса и другие) и устанавливать связи между ними на основании неформализованных документов (тексты, анкеты, карточки и т.д.).

Для обеспечения удаленного взаимодействия пользователей, и обеспечения эффективной совместной работы предназначена подсистема X-Office. В ее состав, в свою очередь, входят такие подсистемы:

- «Корпоративная веб-почта». Обеспечивает работу с корпоративной почтой из любого места по шифрованному каналу связи без необходимости настройки и «следов» на компьютере;
- «Файловое хранилище документов» (рис. 71). Представляет собой удаленное защищенное файловое хранилище, доступное из любой точки только членам закрытой группы. Обеспечивает доступ к личным и корпоративным документам с возможностью совместной работы нескольких пользователей (рис.

72). Хранилище документов предоставляет возможность сквозного поиска по содержанию документов. Разграничение доступа к тексту документа производится согласно профилю доступа или с разрешения автора документа;

- «Переговорная» (рис. 73). Обеспечивает пользователям системы возможность общения внутри закрытой группы в текстовом, голосовом и видео режиме по защищенному протоколу связи. Также доступна возможность совершать звонки на стационарные и мобильные телефоны вне группы с невозможностью определения исходящего абонента.

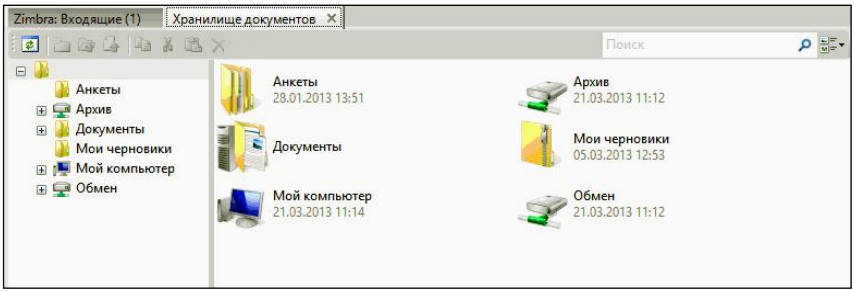


Рис. 71 – Интерфейс работы с хранилищем корпоративных документов

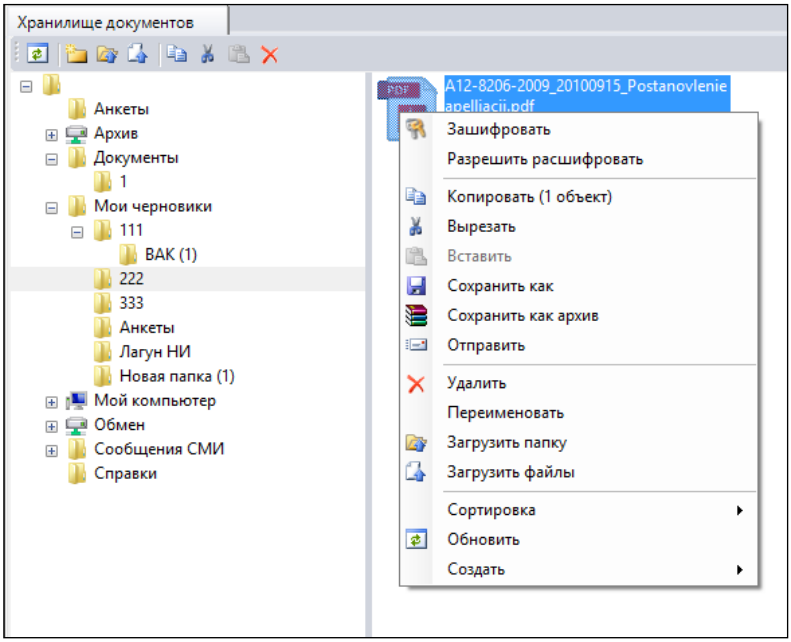


Рис. 72 – Операции, доступные в файловом хранилище

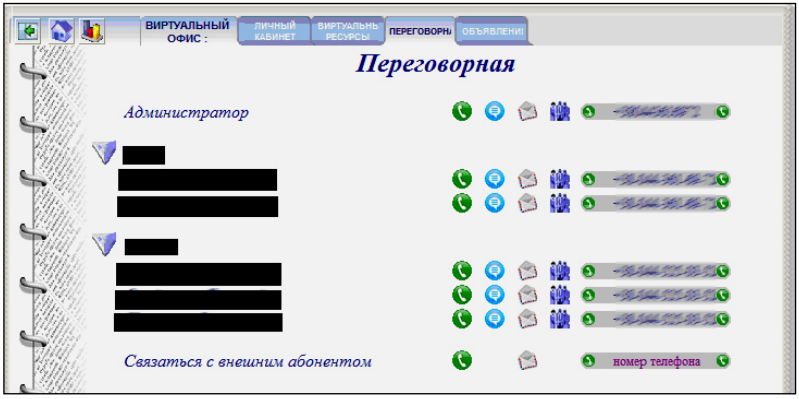


Рис. 73 – Интерфейс подсистемы «Переговорная»

Подистема X-Office обладает рядом особенностей, которые дают ее пользователям существенные преимущества по

сравнению с аналогичными программно-аппаратными комплексами. Авторизация в подсистеме происходит по комбинации отпечатка пальца и пароля, а для коммуникаций используются шифрованные каналы связи и только доверенные сертификаты, что предотвращает возможность анализа трафика на уровне интернет-провайдера или в любой другой точке перехвата. По завершении работы с подсистемой на компьютере пользователя не остается никаких следов ее функционирования.

Файловое хранилище корпоративных документов, которое входит в состав подсистемы X-Office, предоставляет пользователю возможность работы с удаленным облачным хранилищем так же просто, как с папкой на локальном диске его компьютера. Подсистема делает работу с различным источниками (локальный диск, корпоративные документы, файловые хранилища) прозрачной для пользователя. Редактирование офисных документов возможно без установки и настройки дополнительного программного обеспечения. Разграничение доступа к различным файлам осуществляется на основании шаблонов доступа, заданных администратором. В случае наличия особо секретной информации пользователь имеет возможность дополнительно ограничить к ней доступ средствами шифрования прямо из интерфейса программы.

Для проверки благонадежности контрагентов в автоматическом режиме доступна подсистема X-Scoring. Подсистема реализована на основании технологии XML Web service, что обеспечивает прозрачную интеграцию с большинством существующих систем на стороне заказчика. Несмотря на большой объем постоянно пополняющейся информации, на основании которой принимается решение о благонадежности контрагента, система дает ответ менее чем за 3 сек.

Алгоритм принятия решения может гибко настраиваться под потребности конкретного заказчика. Типовой алгоритм проведения проверки и принятия решения о благонадежности контрагента состоит из следующих этапов:

- оценка экономической платежеспособности клиента по предоставленным им сведениям;
- автоматическая проверка по банку данных, нацеленная на проверку соответствия предоставляемых заемщиком сведений и выявление возможных по-

пыток мошенничества со стороны недобросовестных заемщиков;

- детальная проверка контрагента, заявка которого прошла все предыдущие этапы.

Следует отметить, что не всегда полнофункциональные системы конкурентной разведки являются доступными и даже необходимыми, ввиду их стоимостных характеристик или других причин. Вместе с тем, отдельные задачи конкурентной разведки могут быть частично решены вполне доступными средствами. Использование новых подходов, а также открытых, доступных и относительно недорогих информационных источников, позволяет уже сегодня эффективно поддерживать принятие управленческих решений по очень многим, в том числе и стратегическим, направлениям бизнеса.

Технологии конкурентной разведки завтрашнего дня сегодня во многих случаях уже реализованы в виде военных информационных технологий. Так, например, для осуществления разведки на государственном уровне в США еще в 2005 году в рамках Национальной разведки (National Intelligence) была создана специальная структура – Центр открытых источников (Open Source Center). В настоящее время все разведывательные центры США работы с открытыми источниками объединены в эту единую информационную систему. В 2006 г. информационные ресурсы данной системы получили название Intelink-U [Кондратьев, 2010].

Несмотря на то, что материалы в данной системе добываются из общедоступных открытых источников, она предназначена далеко не для всех. Информация из системы распространяется по сетям ограниченного доступа.

В состав системы Intelink-U входят многочисленные базы данных, среди которых:

- база данных CIRC, содержащая свыше 10 млн. статей научно-технической тематики, включая информацию о патентах, стандартах, военном вооружении и военной технике;
- база данных DTED, содержащая большое количество разнообразных карт, полученных от Национального управления геопространственной разведки;

- материалы центров и пунктов информационной службы зарубежного вещания FBIS;
- база данных периодических изданий IC ROSE;
- информационные порталы научно-исследовательских и учебных заведений;
- онлайн-справочники информационной службы Jane's Information Group;
- ресурсы негосударственного информационно-аналитического агентства STRATFOR's, предоставляющего, среди прочего и регулярные обновления по районам развертывания авианосных и экспедиционных ударных групп ВМС США.

Внимания заслуживают и способы наполнения подобных информационных ресурсов. Например, для сбора информации с сайтов всемирной сети Интернет, ее систематизации, перевода и архивирования во Всемирной информационной библиотеке (World Basic Information Library – WBIL), обеспечение функционирования которой возложено на Управление изучения Вооруженных сил (ВС) иностранных государств FMSO Командования учебного и научных исследований по строительству сухопутных войск (СВ) (Training and Doctrine Command – TRADOC), привлекается личный состав резерва СВ и других видов Вооруженных сил.

Для обеспечения потребностей министерства обороны и разведывательного сообщества США создана база данных HARMONY, содержащая библиографические справки по всем имеющимся источникам информации (метаинформацию) о зарубежных странах. База данных HARMONY характеризуется простотой использования, возможностью быстрого поиска необходимых документов, быстрого обмена данными внутри правительственных структур США.

Всемирная информационная библиотека (World Basic Information Library, WBIL) представляет собой специальную программу разведывательного сообщества США, управление которой возложено на отдел изучения вооруженных сил иностранных государств (Foreign Military Studies Office, FMSO) командования учебного и научных исследований по строительству сухопутных войск (Training and Doctrine Command, TRADOC). Персонал с правом доступа к базе данных может осуществлять сбор информации из сети Интернет, ее систе-

матизирование и архивирование в библиотеке WBIL, применяя аналитический инструментарий Pathfinder. Система Pathfinder позволяет за несколько минут проводить анализ 500 тыс. документов из различных баз данных.

Совокупность названных технологий позволяет сотрудникам военной разведки США получать доступ к огромным массивам данных, удовлетворять потребности в разведывательной информации.

Рассмотрим некоторые проекты Агентства национальной безопасности (АНБ) США, ориентированные на сферы добычи данных, аналитики и прогнозирования [Черных, 2013].

Проект тотальной интеграции информационных потоков предполагал создание программных средств, обеспечивающих решение весьма сложной задачи, не решенной по настоящее время в гражданском секторе – интеграцию всех информационных потоков в едином хранилище с одной стороны, и их разделение по специальным критериям, с другой стороны. К настоящему времени, согласно мнению экспертов, задача полностью решена. В военной сфере для извлечения информации из потоков данных, поступающей по каналам АНБ, таким как «Эшелон», «Титан», «Буря», «Эйнштейн», «Интернет Игл» и др. применяется система управления базами данных Prosecutor's Management Information System (PROMIS), разработанная в Inslaw Inc. под руководством Билла Гамильтона (B. Hamilton). Более 30 лет программа, имеющая вначале 570000 строк программного кода, непрерывно совершенствуется собственными разработчиками АНБ.

Программа PROMIS способна одновременно интегрировать неограниченной объем информации, получаемой от неограниченного количества программ и содержащегося в любом количестве баз данных, независимо от их типов, языков, на которых написаны оригинальные программы, архитектуры операционных систем и платформ, откуда извлекается информация. По-видимому, аналогов PROMIS в мире не существует.

Другим важным направлением является добыча знаний в реальном режиме времени. Сегодня, согласно имеющейся информации, машины могут извлекать знания примерно о 40 млн. сущностей (объекты, субъекты, события и др.) и более чем 2,5 трлн. параметров.

По заказу DARPA компания Raytheon создала самоорганизующуюся базу знаний, которая позволяет автоматически составлять досье на граждан и организации, собирая информацию из открытых источников. С конца 2011 г. головными разработчиками таких баз знаний стали IBM и Recorded Future. Уже сегодня по заказу отдельных родов войск США им удалось создать эффективные системы раннего предупреждения кризисов.

Очень большие надежды связываются с проектом по автоматизированному выявлению аномальных процессов, протекающих в различных масштабах. Источниками информации для программы являются как обычно веб 1 и веб 2, а также анализ потокового видео, финансовых транзакций и т.п.

Большое внимание АНБ и американское разведывательное сообщество уделяли и уделяют проекту по анализу информации и прогнозированию в реальном масштабе времени. Наиболее известной реализованной автоматизированной системой (с участием человека-эксперта) в рамках этого проекта является Palantir, разработанная компанией Palantir Technologies, которая предназначена для анализа и визуализации данных. Система обеспечивает сбор потоков данных со всех доступных информационных каналов обо всех регистрируемых событиях, касающихся людей: банковские транзакции, транзакции по кредитным карточкам, телефонные звонки, электронная почта, информация с камер видеонаблюдения, информация о транзакциях во всех федеральных и муниципальных базах данных и т.п. В каждом потоке данных средствами интеллектуального анализа данных выявляются необычные события, вероятность которых мала, и события из наперед заданных «тревожных списков». Затем происходит объединение и увязка информации о необычных событиях из разных баз данных. И в результате, если расчетная вероятность всего комплекса связанных необычных событий окажется ниже некоего заданного порога вероятности, выдается сигнал тревоги, указывающий на конкретное лицо, с которым связан весь комплекс событий.

Программа Palantir постепенно находит применение и в гражданском секторе. Так, представитель банка JPMorgan Chase Гай Чарелло (Guy Chiarello) говорит, что программы

Palantir помогают банку выявлять мошенников еще до того, как случилось преступление. Контракт с Palantir, по его словам, – это «лучшая сделка за последнее время».

При этом основные инвесторы проекта Алекс Карп (Alex Karp) и Питер Тиль (Peter Thiel) говорят, что величайшая проблема, которую им удалось решить с помощью своих программ, – это возможность борьбы с терроризмом и насилием при сохранении гражданских свобод. Задействованная в их работе информация засекречена, и пользователи имеют доступ лишь к отдельным ее фрагментам.

Для качественного проведения конкурентной разведки методами анализа текстов из сети Интернет необходимо сформулировать цели, построить базы данных для наблюдений и проведения исследований, сформулировать запросы. Заметим, что не следует ограничиваться одной информационно-поисковой системой даже для анализа такой информации, как интернет-ресурсы. Рекомендуем использовать лучшие глобальные и специальные информационно-поисковые системы, такие как Google (www.google.com), Yahoo! (www.yahoo.com) или Microsoft Bing (www.bing.com). Для специальных потребностей рекомендуется также использовать законодательные, адресно-номенклатурные, ценовые базы данных, доступные как из сети Интернет, так и в локальных версиях.

Покажем, как формируются запросы, относящиеся к конкурентной проблематике, на примере поисковых предписаний к системе контент-мониторинга InfoStream (www.infostream.ua, Рис. 74).

Обычно поиск информации о компании или персоне всегда начинается с указания различных способов написания названия компании или Ф.И.О. персоны. Порой поиска в оперативных и ретроспективных данных по таким «примитивным» запросам вполне достаточно, однако задача усложняется, если необходимо исследовать состояние отдельной отрасли, отдельного региона, или даже целой страны. В таких случаях в соответствии с проблематикой строятся запросы, которые затем итеративно уточняются.

В качестве примера приведем ряд понятий, а затем поставим им в соответствие фрагменты запросов и рассмотрим фрагменты текстов, публикуемые различными источниками,

которые затем можно использовать при построении разного рода аналитических справок.

Рис. 74 – Веб-сайт системы контент-мониторинга InfoStream

После нахождения документов, содержащих упоминания анализируемых фирм или брендов, можно путем уточнения запросов выяснить некоторые важные характеристики, относящиеся к деятельности этих компаний. В качестве примера ниже приведены уточняющие запросы, относящиеся к финансовому положению компаний, упоминаемых в веб-пространстве:

Уставн~капитал~/2/грн
Уставн~капитал~/2/долл
Уставн~фонд~/2/грн
Уставн~фонд~/2/долл
Принадлежит~/2/акций

В первых двух запросах подразумевается нахождение документов, в которые входят фрагменты, содержащие словосочетания «уставной капитал» или «уставный фонд», с указанием значения в долларах или гривнях («~/2/»; на языке за-

просов это означает расстояние в 2 или менее слов между выражениями).

В результате поиска получены текстовые документы, содержащие такие фрагменты:

Антимонопольный комитет согласовал приобретение компанией Luregio Limited "Проминвестбанка" через покупку ООО "Финансовая компания "Фортифай". В соответствии с информацией в Едином госреестре юрлиц и физлиц- предпринимателей, Luregio Limited 28 августа этого года зарегистрировало ООО "Лурежио инвест" с **уставным капиталом 350 млн грн**, основными видами деятельности которого указаны консультирование и управление, предоставление финансовых услуг и других вспомогательных коммерческих услуг. Руководителем предприятия назначена Людмила Назаренко, которая была замдиректора по юридическим вопросам ООО "Группа ТАС" и членом набсовета "Универсал банка", принадлежащих Тигипко.

Fin.org.ua 2021.02.05 18:15

Министерство инфраструктуры определило первого участника экспериментального проекта по допуску частных локомотивов к работе отдельным маршрутам на железнодорожных путях общего пользования. Им стало ООО "Украинская локомотивостроительная компания". ООО "Украинская локомотивная компания" с **уставным капиталом 3 тыс. грн** основано в Киеве в ноябре 2016 года. Его бенефициаром через ООО "Инвестиционная компания "Евразия" указан директор Вячеслав Якубовский.

Бизнес Цензор 2021.02.05 18:02

Комиссия по регулированию игорного бизнеса и лотерей 2 февраля выдала ООО "СПЕЙСИКС" (ТМ "Космолот") лицензию на осуществление деятельности по организации и проведению азартных игр казино в сети Интернет. По данным портала YouControl, ООО "СПЕЙСИКС" зарегистрировано 28 мая 2020 в Киеве, рамер **уставного капитала - 30 млн грн**. Основатель и руководитель - Сергей Потапов. Основной вид деятельности компании - организация азартных игр.

IPress.ua 2021.02.03 19:37

"Центрэнерго" эксплуатирует 23 блока (18 - пылеугольные и пять - газомазутные) на Углегорской, Змиевской и Трипольской ТЭС суммарной установленной мощностью 7660 МВт. Государству **принадлежит 78,3% акций** общества.

Интерфакс-Україна 2021.02.04 21:01

Глава НАК Андрей Коболев заявлял, что с юридической точки зрения механизм разделения "Укрнафты" сложен, и будет необходима помощь со стороны органов государственной власти, чтобы его реализовать. "Нафтогаз Украины" владеет 50%+1 акцией "Укрнафты", а группе компаний, связанных с Игорем Коломойским, **принадлежит около 42% акций**. На балансе компании находится 25 буровых установок, 1891 нефтяных и 162 газовых скважин, компании принадлежит 537 АЗС.

Интерфакс-Україна 2021.02.04 21:01

Информация о слияниях и приобретениях в той или иной сфере бизнеса, позволяющая следить за экспансией конкурентов в новые рыночные ниши, может быть получена в результате отработки таких уточняющих запросов:

Приобр~/2/акций
приобр~/2/пакет~акций
продаж~/2/пакет~акций
(слияние~компаний) & (акций,активов)

Выполнение этих уточняющих запросов позволяет получить документы, содержащие, например:

ПАО "Проминвестбанк" создано в 1992 году. Российский ВЭБ стал владельцем "Проминвестбанка" в 2008 году, **приобретя пакет его акций** в размере 99,7726%. Свои инвестиции в развитие украинского дочернего банка ВЭБ оценивал в $2,7 млрд.

UaProm 2021.02.04 17:49

Нынешний владелец итальянской команды Габриэле Вольпи готов рассмотреть вариант продажи. Недавн Пейс **приобрел**

84 процента акций клуба АПЛ "Бернли". За эту сделку он заплатил 170 миллионов евро.

Dynamomania.com 2021.02.04 19:34

КИЕВ. 2 февраля. УНН. Президент Владимир Зеленский не приводит законных причин для блокирования доступа китайских акционеров к управлению "Мотор Сич", поэтому его действия являются политически мотивированными. На это обращает внимание в своей публикации Виктор Суслов, экс-министр экономики, передает УНН. "С точки зрения законодательства, не ясно, что значит "президент намерен не допустить покупки контрольного пакета акций", - комментирует эксперт личное вмешательство президента Украины в законную сделку по **продаже контрольного пакета акций** "Мотор Сич". - Ведь при этом Зеленский не говорит, что китайцы поступили не по закону или нарушили какие-то нормы".

FinOboz 2021.02.02 22:34

В 2007 году компания Wuhan Iron and Steel, через девять лет ставшая одним из основателей группы Baowu, **приобрела 48,81% акций** Kunming Iron and Steel Joint Stock - крупнейшего металлургического предприятия Kunsteel. Ранее сообщалось о намерении Baowu присоединить к себе еще одну китайскую металлургическую компанию Shandong Steel. Благодаря этим поглощениям производственные мощности группы превысят 150 млн. т в год.

UaProm 2021.02.04 12:49

Для выявления публикаций об изменении финансового состояния и банкротства можно использовать такие уточняющие запросы:

```
выпуск~/2/акц
(увелич~уставн) & (фонд,капитал)
повыс~/1/дол~/2/акц
объяв~/2/банкротств
```

Отработка подобных запросов позволила найти такие документы:

В биржевом списке ПФТС находились 119 выпусков внутренних государственных облигаций, 12 выпусков внешних государственных облигаций, 65 выпусков корпоративных облигаций, 101 **выпуск акций** предприятий, 23 выпуска ценных бумаг институтов совместного инвестирования, 15 выпусков муниципальных облигаций, один государственный дериватив, одна облигация иностранного государства, два выпуска облигаций иностранного эмитента, три выпуска акций иностранных эмитентов и три выпуска ценных бумаг институтов совместного инвестирования иностранных эмитентов.

E-Finance 2021.02.05 09:30

Городской совет Одессы **увеличил уставный капитал** КП "Одесгорэлектротранс". Он вырос на 45 млн грн - до 108,5 млн грн. Средства, добаленные в уставный капитал КП, пойдут на очередной платеж (нужно внести до 25 марта) по кредитному договору с Европейским банком реконструкции и развития. За счет кредита Одесса купила 47 новых троллейбусов.

Пора говорить 2021.02.05 13:36

После аудита Wirecard признала, что записала в актив 1,9 млрд евро с несуществующих банковских счетов. Курс ее акций на бирже рухнул, а сама она **объявила о банкротстве.** Федеральное ведомство по надзору за финансовым сектором назвало ситуацию "катастрофой" и "позором". Международное рейтинговое агентство Moody's отозвало рейтинг Wirecard.

Goodnews.ua 2021.01.30 06:42

Сеть "Фуршет" **объявила о банкротстве**. Хозяйственный суд Днепропетровской области начал процесс оформления банкротства компании "Ритейл Центр", являющейся владельцем сети "Фуршет". Компании-поставщики розничной сети "Фуршета" получили письмо с подписью директора "Ритейл Центр" В. Купченко о том, что суд начал процедуру банкротства компании, в связи с чем вводится мораторий на выплаты кредиторам и назначен распорядитель имущества.

Матриця свободи 2021.01.27 19:03

Методы контент-мониторинга – это адаптация классических методов контент-анализа к условиям динамических информационных массивов, например потоков информации из сети Интернет.

Типичная задача контент-мониторинга – построение диаграмм динамики появления понятий по времени.

Рассмотрим, как в системе InfoStream отслеживались кризисные явления на рынке продуктов питания в Украине в июне 2011 года. Для этого был составлен запрос ***«кризис & гречка & Украина»***, который был введен через веб-интерфейс системы. В специальном режиме «Динамика» была получена соответствующая диаграмма появления понятия (рис. 75).

Рис. 75 – Динамика появления понятия

На приведенной диаграмме видно, что массовое появление сообщений о кризисных явлениях произошло 12-го января 2011 г. (в то время, как сами цены на гречневую крупу резко возросли лишь в середине марта).

Безусловно, оперативное получение такого типа данных должно было помочь аналитикам при построении краткосрочных прогнозов.

Аналогично можно проводить мониторинг финансового рынка. К примеру, простой запрос «паден~курс~гривн», относящийся к фрагменту информационного потока за период с января по март 2015 года выдал диаграмму, свидетельствующую о динамике падения курса украинской валюты (рис.

76). Как видим, пик публикаций приходится на 27 февраля 2015 г., когда, в частности, произошло «локальное» обесценивание гривны до 44 грн/долл.

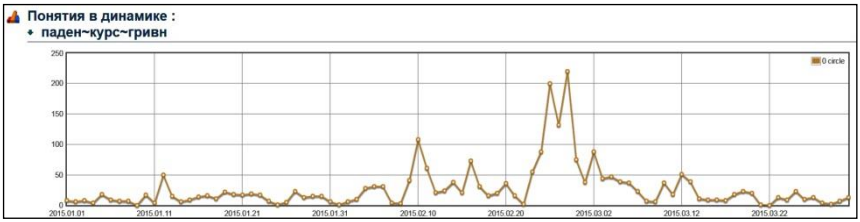


Рис. 76 – «Падение курса гривни» в динамике

3. Источники информации

В информационно-аналитической работе важное значение имеет возможность доступа к источникам данных, информации и знаний. При этом главной проблемой является нахождение содержательных и надежных источников из всех общедоступных. Когда такие источники найдены, включаются механизмы превращения данных в знания, для чего применяются соответствующие технологии. Под данными обычно понимают «сырые», необработанные сведения, основанные на фактах. Это могут быть статистические данные, факты из биографий ключевых персон или, например, сведения об отчетности отдельных компаний. Информация представляет собой уже определенным образом обработанные и проанализированные данные. Конечным же информационным продуктом любой аналитической работы являются знания – синтезированные выводы, рекомендации для принятия решений.

Информация, как было указано выше, может быть получена из официальных, открытых источников, СМИ, объявлений, рекламы, фирменных, банковских, правительственных отчетов, баз данных, от экспертов путем анализа или специальной обработки данных, текстов.

Ниже приведен подробный перечень видов информационных источников, которые чаще всего используются при конкурентной разведке [Нежданов, 2009].

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

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

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

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

5. Патенты, авторские свидетельства компании и ее сотрудников. Для задач конкурентной разведки интересно их содержание, направленность, списки соавторов. Информация размещается на соответствующих сайтах. Для Украины: https://ukrpatent.org/; Google Patents: https://patents.google.com/; Евразийское патентное ведомство: www.eapo.org. Патентование возможно в любой стране, предпочтительные варианты – страна регистрации организации, страна ведения бизнеса, кроме того США, Евросоюз, Россия, Япония и Китай.

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

7. Активность компании на рынке слияний и поглощений (M&A). Информация о том, какие организации поглощаются, планируют поглотить или ведут переговоры о поглощении. Информацию можно получить в Антимонопольном комитете (АМК) Украины, аналогичных ведомствах других стран, по новостным сообщениям на веб-ресурсах посвященных M&A.

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

9. Курсы повышения квалификации, обучение персонала – указание на приоритеты в развитии компании. Интерес представляет то, чему обучают, каких специалистов приглашают для обучения, какие требования выдвигают при привлечении обучающих, какие сроки обучения, какое количество персонала обучается.

10. Благодарности и награды компании и ее сотрудников.

11. Участие в мероприятиях (выставки, конференции, круглые столы, презентации). Выяснение, в каких мероприятиях участвуют компании, их направленность, круг участников.

12. Участие в организациях (союзы, ассоциации, конфедерации и т.п.) – информация о том, в каких объединениях участвует компания, как активно участвует, что получает от участия, на что рассчитывает, как использует.

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

Деятельность конкурентной разведки основана на использовании только легитимных источников информации, которых вполне достаточно для принятия управленческих решений в сфере бизнеса, необходимо лишь провести некоторую информационно-аналитическую обработку имеющихся открытых данных. Среди таких источников информации можно назвать: данные статистики, материалы с веб-сайтов, социальных сетей, СМИ, отраслевых отчетов и т.д.

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

У конкурентной разведки в последние десятилетия появился и развился до невиданных ранее масштабов новый информационный источник – веб-пространство сети Интернет. Сегодня по оценкам экспертов Интернет по количеству информации находится на первом месте, опережая СМИ, отраслевые издания и получаемые от коллег новости, специальные обзоры, закрытые базы данных. При этом в открытых источниках и специализированных базах данных, доступных в Интернет, содержится большая часть информации, необходимой для проведения конкурентной разведки, однако остается открытым вопрос ее нахождения и эффективного использования. Последние исследования информационного веб-пространства показали, что доступный через традиционные информационно-поисковые системы триллион веб-страниц – это лишь «поверхностная видимая часть айсберга». Около 40 % всей информации в Интернете доступно бесплатно. Навигацию по данному информационному пространству обеспечивают более миллиона поисковых систем и каталогов, но и они охватывают лишь малую часть информационных ресурсов. Скрытых и невидимых (deep, invisible) ресурсов сети Интернет значительно больше – это, прежде всего динамически-генерируемые страницы, файлы разнообразных форматов, информация из многочисленных баз данных. К «скрытому» веб можно отнести и такие сети, как BitTorrent, DirectConnect, EMule, Napster и др.

Сегодня для конкурентной разведки основными источниками информации служат Интернет, пресса, а также открытые базы данных. Очень популярны среди специалистов по конкурентной разведке базы данных государственных и статистических органов, торгово-промышленных палат, органов приватизации и т.д. Большую пользу приносят и отдельные доступные базы данных других органов власти. В последнее время все более популярны базы данных на основе архивов СМИ, в том числе и сетевых. В России, например, большой популярностью пользуется крупнейшая архивная база данных СМИ службы «Интегрум» (integrum.ru), содержащая несколько сотен миллионов документов. С помощью другой российской базы данных «Лабиринт» (labyrinth.ru), составленной на основе публикаций ведущих бизнес-изданий, можно полу-

чить обширную информацию о конкретных персонах, организациях и компаниях.

Традиционно конкурентная разведка опирается на следующие источники информации, как опубликованные документы открытого доступа, которые содержат обзоры товарного рынка, информацию о новых технологиях, создании партнерств, слияниях и приобретениях, объявлениях о рабочих вакансиях, о выставках и конференциях, и т.п. Широко используются сведения, находящиеся в документах, уже имеющихся в компаниях, ведущих конкурентную разведку, результаты маркетинговых исследований, информация, полученная на конференциях, при общении с клиентами и коллегами. Большая часть этих данных попадает в сетевую прессу, пресс-релизы или публикуются на корпоративных веб-сайтах.

Поэтому в последнее время большую популярность получают базы данных на основе архивов масс-медиа, в том числе (и преимущественно) сетевых.

3.1. Веб-сайты

Веб-пространство, основанное на физической инфраструктуре сети Интернет и протоколе передачи данных HTTP, объединяет сотни миллионов веб-серверов, подключенных к сети Интернет (Рис. 77). В начале существования веб-пространства на небольшом количестве веб-сайтов публиковалась информация отдельных авторов для относительно большого количества посетителей. Сегодня ситуация резко изменилась, произошел переход к веб второго поколения. Сами посетители веб-сайтов активно участвуют в создании контента, что привело к резкому росту объемов информации и динамики веб.

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

Server Survey (news.netcraft.com, Рис. 77), в настоящее время количество адресов веб-сайтов превышает 1 198 млн. (из низ около 200 млн. активных).

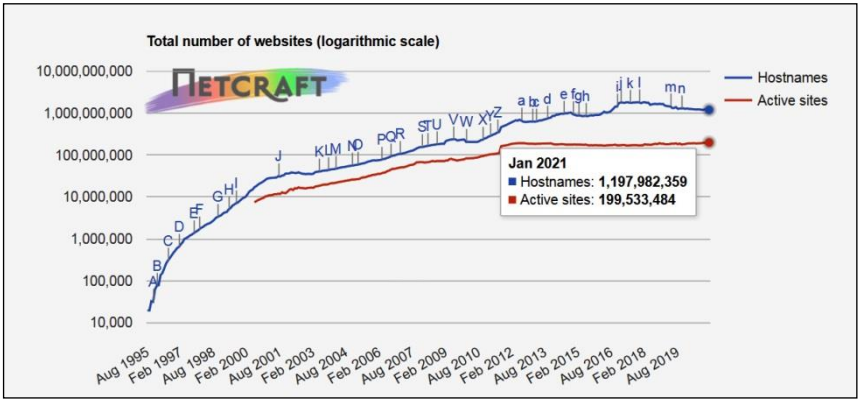


Рис. 77 – Динамика роста количества веб-серверов по логарифмической шкале (Netcraft, январь 2021 года)

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

Возможности доступа к интернет-ресурсам, привлекающим своей открытостью, объемами и содержательной многогранностью, на первый взгляд кажутся безграничными. Однако важные события в различных областях свидетельствуют об обратном. Именно в кризисных ситуациях Интернет довольно часто подводит. Существует множество проблем – от перегруженности сетевой инфраструктуры – до вирусных атак, уязвимостей и отказов в обслуживании отдельных веб-серверов. Целый ряд проблем порожден также объемами,

разнообразием представления и динамикой контентного сегмента информационного пространства.

Несмотря на такие качества, как открытость и доступность, существующую инфраструктуру веб-пространства нельзя признать надежной и достоверной. Назовем еще несколько проблем, присущих веб-пространству:

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

Известно, что сегодня существуют технологии интеграции контента, которые позволяют частично решать названные проблемы, обеспечивая эффективный поиск и навигацию в веб-пространстве, мониторинг и агрегацию открытых веб-ресурсов. Для профессионального поиска и агрегации информации из веб-пространства используются специализированное программное обеспечение, информационно-поисковые системы и сервисы. Приведем некоторые примеры программных продуктов и сервисов:

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

Newprosoft Web Content Extractor (www.newprosoft.com) – программа сканирования и извлечения данных из веб-сайтов, котоая автоматизирует процесс извлечения данных и позволит сохранять извлеченные данные в выбранном пользователем формате. Web Content Extractor - мощная и простая в использовании программа для очистки веб-страниц. Он поз-

воляет извлекать определенные данные, изображения и файлы с любого веб-сайта. Процесс извлечения веб-данных полностью автоматический. Web Content Extractor имеет удобный интерфейс, правила сканирования и шаблон извлечения обеспечивают эффективное и точное извлечение данных.

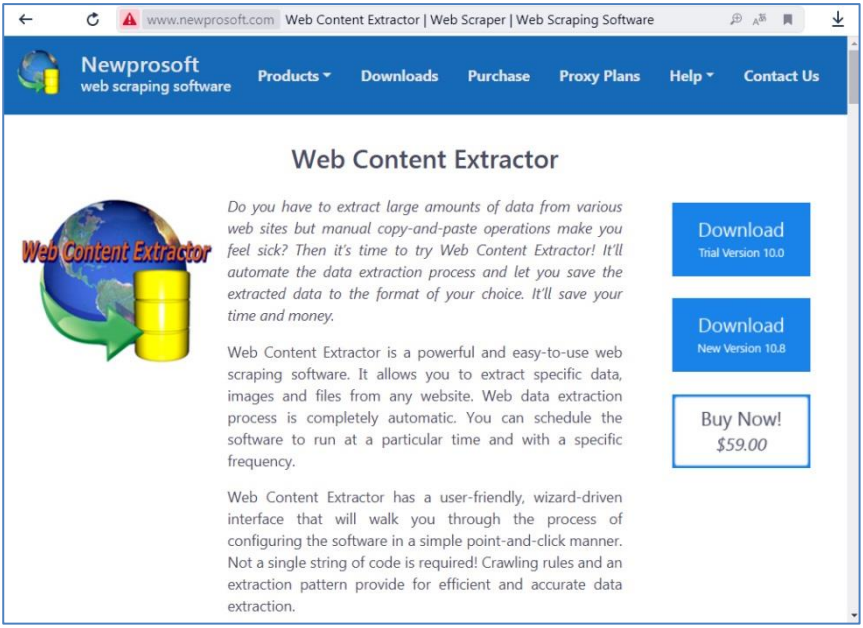


Рис. 78 – Фрагмент веб-сайта Newprosoft Web Content Extractor

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

В качестве сервисных решений можно назвать:

WatchThatPage (watchthatpage.com) – бесплатный сервис, позволяющий автоматически собирать новую информацию с веб-ресурсов, поставленных на мониторинг. Пользователь выбирает, какие страницы отслеживать, а

WatchThatPage определяет, какие страницы были изменены, и предоставляет ему весь новый контент. Новая информация предоставляется по электронной почте.

Newspaper Map (newspapermap.com) – сервис, объединяющий геолокацию и информационно-поисковую систему по медиа-ресурсам. При решении задач конкурентной разведки пользователь может выбрать интересующий его регион, язык, список онлайн версий газет и журналов, непосредственно выходить на документы. Сервис поддерживает русский язык, имеет удобный интерфейс (рис. 79).

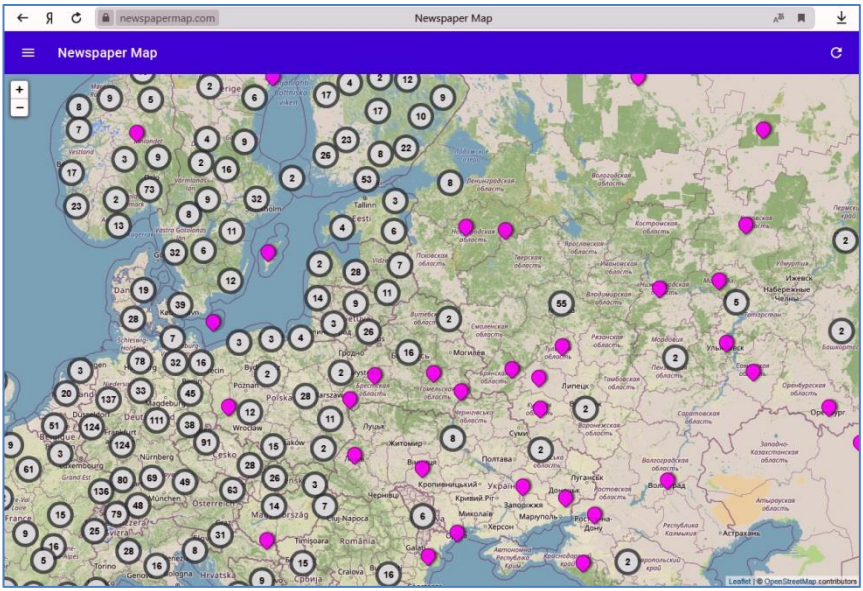


Рис. 79 – Фрагмент географического агрегатора новостей Newspaper Map

WebSvodka (websvodka.ru) – это инструмент автоматического контроля изменений интернет-страниц, который позволяет оперативно узнавать: о появлении новых вакансий и объявлений; об изменении прайс-листов; о новых приказах, распоряжениях, тендерах, конкурсах; контролировать упоминания интересующих вас слов. Пользователь зада-

ет интересующую его тему или страницу, а WebSvodka регулярно сообщает ему обо всех событиях по этой теме. Для повышения надежности и скорости работы модули загрузки контента, анализа страниц и хранения результатов функционируют параллельно и размещены на разных серверах.

InfoStream (www.infostream.ua) – сервис контент-мониторинга веб-ресурсов, предоставляющий доступ в поисковом режиме к информации из 10000 источников, классификацию информации, экстрагирование понятий (персон, компаний, топонимов), формирование сюжетных цепочек, оценку тональности сообщений, анализ динамики публикаций по определенным объектам. В базах данных системы хранится свыше 500 млн. новостных документов за 25 лет.

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

3.2. Социальные сети, блоги

Термин «социальная сеть» обозначает сосредоточение социальных объектов, которые можно рассматривать как сеть (или граф), узлы которой – объекты, а связи – социальные отношения. Этот термин был введен в 1954 году социологом из «Манчестерской школы» Дж. Барнсом (J. Barnes) в работе «Классы и сборы в норвежском островном приходе». Во второй половине XX столетия понятие «социальная сеть» стало популярным у западных исследователей, при этом как узлы социальных сетей стали рассматривать не только представителей социума, но и другие объекты, которым присущий социальные связи. Сегодня термин «социальная сеть» обозначает понятие, оказавшееся шире своего социального аспекта, оно включает, например, многие информационные сети, в том числе и WWW. Рассматривают не только статистические, но и динамические сети, для понимания структуры которых необходим учет принципов их эволюции.

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

1) предоставление пользователям широкого спектра возможностей для обмена информацией;

2) создание профилей пользователей, в которых требуется указать некоторое количество персональной информации;

3) друзьями в социальной сети становятся преимущественно не виртуальные, а реальные друзья.

Веб-ресурс социальной сети предоставляет возможности:

1) активного общения;

2) создания публичного или закрытого профиля (*Profile*) пользователя, содержащего персональные данные;

3) организации и ведения пользователем списка других пользователей, с которыми у него имеются некоторые социальные отношения;

4) просмотра связей между пользователями внутри социальной сети;

5) образования групп пользователей по интересам;

6) управления содержимым в рамках своего профиля;

7) синдикации контента;

8) подключения различных приложений.

Социальные медиа представляют собой совокупность онлайн-сервисов и интернет-приложений, которые позволяют пользователям общаться друг с другом в том числе, и в режиме реального времени. При этом пользователи могут обмениваться между собой мнениями, новостями, информацией, в том числе и мультимедийной.

Социальные медиа базируются на идеологической и технологической базе веб 2.0, позволяющих создание и обмен контентом, созданным самими пользователями (User-Generated Content), в отличие от предшествующей концепции веба, предполагающей, как и в случае традиционных СМИ, централизованное создание контента, поставляемого пользователям-читателям.

Очевидно, социальные медиа являются самым ценным источником информации для конкурентной разведки, предо-

ставляя абсолютно на легальных условиях разностороннюю информацию о людях, событиях, компаниях, брендах, продуктах. Получившие в последнее время широкое распространение такие явления, как информационные операции, активное информационное противодействие в рамках конкурентной борьбы, сетевая мобилизация, во многих случаях базируется на манипулировании данными именно в социальных медиа.

Выделяют семь разновидностей социальных медиа, это социальные сети; блоги; форумы; сайты отзывов; серверы фото- и видеохостинга; виртуальные службы знакомств и геосоциальные сети. Следует отметить, что четкие границы между этими разновидностями размыты.

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

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

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

Веб-сайты отзывов создаются с целью повышения эффективности и качества предоставляемых (не обязательно в интернет-среде) услуг и товаров. Пользователи, посещая веб-

сайты отзывов, оставляют там свои сообщения, участвуют в анкетированиях, формируют мнения о той или иной услуге или товаре.

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

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

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

3.2.1. Основные социальные сети

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

- Facebook;
- Twitter;
- LinkedIn;
- Sina Weibo;
- Youtube;
- Telegram;
- Medium;

- Reddit;
- Livejournal.

Facebook (www.facebook.com) – крупнейшая социальная сеть, основанная в 2004 году М. Цукербергом и его компаньонами. Начиная с сентября 2006 года социальная сеть доступна для пользователей сети Интернет. На июнь 2017 года аудитория *Facebook* составила 2 миллиарда пользователей. Суточная активная аудитория в марте составила 720 миллионов человек. Около 500 млн. человек в месяц используют мобильные приложения *Facebook*. Каждый день в социальной сети пользователи оставляют 6 миллиардов «лайков» и комментариев и публикуют 300 миллионов фотографий. На сайте зафиксировано 200 миллиардов «дружеских связей». Ежемесячное количество просмотров страниц Facebook превышает 1 триллион.

Twitter (twitter.com) – сервис, позволяющий пользователям отправлять короткие текстовые заметки (до 140 символов), используя веб-интерфейс, SMS, средства мгновенного обмена сообщениями или сторонние программы-клиенты. Созданный Джеком. Владельцем системы *Twitter* является компания *Twitter Inc.*, главный офис которой находится в Сан-Франциско. По состоянию на 1 января 2011 года сервис насчитывает более 200 млн пользователей. 100 млн пользователей проявляют активность хотя бы раз в месяц, из них 50 миллионов пользуются *Twitter* ежедневно. 55 % пользуются Twitter на мобильных гаджетах, около 400 миллионов уникальных посещений получает за месяц непосредственно сайт twitter.com. Особенностью *Twitter* является публичная доступность размещённых сообщений; это называется микроблоггингом.

LinkedIn (www.linkedin.com) – социальная сеть для поиска и установления деловых контактов. Социальная сеть *LinkedIn* была основана Ридом Хоффманом в декабре 2002 года, запущена в мае 2003 года. 13 июня 2016 года Microsoft объявила о приобретении *LinkedIn* по цене 196 долларов за акцию (общая цена сделки — 26,2 миллиарда долларов), что является крупнейшим приобретением Microsoft на сегодняшний день. В конце июля 2020 года *LinkedIn* объявила об увольнении 960 сотрудников; сокращения были вызваны

последствиями глобальной пандемии Covid-19. Эта социальная сеть предоставляет возможность зарегистрированным пользователям создавать и поддерживать список деловых контактов. Контакты могут быть приглашены как из сайта, так и извне, однако *LinkedIn* требует предварительного знакомства с контактами. Список контактов *LinkedIn* может использоваться для расширения связей, поиска компаний, людей и групп по интересам, публикации резюме и поиска работы рекомендовать пользователей, публиковать вакансии, создавать группы по интересам. Социальная сеть *LinkedIn* также позволяет публиковать информацию о деловых поездках и конференциях. На 2020 год общее число пользователей *LinkedIn* достигло 675 миллионов, из них 310 миллионов активных.

Sina Weibo (кит. 新浪微博, http://weibo.com) –сервис микроблогов, запущенный компанией Sina Corp 14 августа 2009 года, один из самых популярных интернет-сервисов (платформ социальных сетей) в Китае и в мире. В начале 2018 года он превысил рыночную оценку в 30 миллиардов долларов США. По состоянию на февраль 2013 года число пользователей сервиса составляет более 500 миллионов. В июне 2020 года Sina Weibo достигла 523 миллионов активных пользователей в месяц.

Youtube (youtube.com) – видеохостинг, предоставляющий пользователям услуги хранения, доставки и показа видео. Сервис, созданный в феврале 2005 года тремя бывшими сотрудниками PayPal - Чадом Херли, Стивом Ченом и Джаведом Каримом, - был куплен Google в ноябре 2006 года за 1,65 миллиарда долларов США. Согласно рейтингу Alexa в Интернете, *YouTube* является вторым по посещаемости веб-сайтом после Google Search. Пользователи могут загружать, просматривать, оценивать, комментировать, добавлять в избранное и делиться теми или иными видеозаписями. По состоянию на 2019 на *YouTube* загружают около 300 часов видео каждую минуту, а количество ежедневных просмотров видео достигло 5 млрд.

Telegram (telegram.org) – кроссплатформенный мессенджер, позволяющий обмениваться сообщениями и медиафай-

лами многих форматов. Пользователи могут отправлять сообщения и обмениваться фотографиями, стикерами, голосовыми и видео сообщениями, файлами любого типа, а также делать аудио- и видеозвонки. Количество ежемесячных активных пользователей сервиса, по состоянию на январь 2021 года, составляет около 500 млн человек. По состоянию на март 2020 года официальные клиенты для *Telegram* включают в себя:

- Мобильные приложения для Android и iOS/iPadOS;
- Настольные приложения для Windows, Linux и macOS;
- Веб-приложение, веб-приложения для Chrome app, веб-приложение для React.

С 28 января 2021 года в *Telegram* появилась возможность импортировать чаты и историю переписки из других мессенджеров в том числе и *WhatsApp*.

Medium (medium.com) – платформа для социальной журналистики. Сервис запущен в августе 2012 года сооснователями *Twitter* – Эваном Уильямсом и Бизом Стоуном. Уильямс, ранее соучредитель Blogger и Twitter, первоначально разработал *Medium* как средство публикации писем и документов, длина которых превышает максимум 140 символов *Twitter* (теперь 280 символов). Среди 75 сотрудников компании 15 авторов и редакторов. Платформа выпускает издания Matter, Cuepoint, Backchannel, Re: form, Vantage и The Nib. По состоянию на май 2017 года у *Medium* было 60 миллионов уникальных читателей в месяц.

Reddit (reddit.com) – социальный новостной сайт, на котором зарегистрированные пользователи могут размещать ссылки на какую-либо понравившуюся информацию в интернете. Как и многие другие подобные сайты, *Reddit* — один из наиболее популярных сайтов в мире, занимает 19-е место по посещаемости по данным Alexa Internet. *Reddit* был основан 23 июня 2005 года выпускниками Виргинского университета Стивом Хаффманом и Алексисом Оганяном. Оценочная стоимость *Reddit* составляет $6 млрд. В 2019 году ежемесячно насчитывалось около 430 миллионов пользователей *Reddit*, известных как «реддиторы».

LiveJournal, LJ (www.livejour-nal.com) – платформа для ведения онлайн-дневников (блогов), созданная в 1999 г. американским программистом Брэдом Фицпатриком. *LiveJournal* предоставляет пользователям возможность публиковать свои и комментировать чужие записи, вести коллективные блоги («сообщества»), добавлять в друзья других пользователей и следить за их записями в «ленте друзей». До конца декабря 2016 года серверы *LiveJournal* находились в США и система принадлежала американской компании LiveJournal, Inc., но с декабря 2016 года *LiveJournal* размещен на серверах российской компании Rambler&Co. Среди опций «Живого Журнала» следует выделить: разные типы записей и возможности их комментирования; указание расширенных сведений о пользователе; друзья и лента друзей; картинки пользователей; функции безопасности аккаунта. на конец 2012 года, в *LiveJournal* было зарегистрировано более 40 млн пользователей, из них 368 805 активных.

3.2.2. Мониторинг социальных сетей

Мониторинг социальных медиа – важнейший этап для успешного развития бизнеса, продвижения в Интернет, конкурентной разведки. С помощью социальных медиа можно узнать наиболее полную информацию об аудитории товара или услуги, ее мнении о работе компании.

Приведем пример нескольких сервисов для эффективного мониторинга социальных медиа, сосредоточив внимание на наиболее доступных:

Socialmention (www.socialmention.com) – платформа бесплатного поиска и анализа информации в социальных сетях. Слоган – «Поиск и анализ в социальных сетях в реальном времени». Система ищет упоминания в выбранных сетях или во всех сетях сразу. Предоставляет анализ тональности упоминаний, связанные ключевые слова, популярные источники и многое другое. Охват системы – более 100 социальных медиа, включая социальные сети, социальные закладки, блоги, форумы и многое другое.

Hootsuite (hootsuite.com, seesmic.com) – сервис мониторинга социальных медиа. Слоган сервиса – «Социальные сети - это ваша суперсила». Поддерживает мониторинг таких ресурсов, как: Twitter, Facebook, LinkedIn, Chatter, Ping.fm. Есть приложения как для веб, так и для персонального компьютера, iPhone, Android, Windows Mobile. Сервис HootSuite является сертифицированным партнёром Twitter. Обеспечивает постинг (posting) по расписанию, возможность отслеживать сообщения по ключевым словам и упоминаниям. Система HootSuite также предоставляет полноценную интеграцию с Facebook.

YouScan (www.youscan.io) – система мониторинга русскоязычных социальных медиа. Слоган – «Используйте силу соцмедиа для принятия верных решений». Система YouScan отслеживает упоминания в блогах, форумах, социальных сетях (Facebook, ВКонтакте), Twitter, YouTube, и предоставляет результаты мониторинга в аналитическом интерфейсе с функциями одновременной работы нескольких сотрудников. Представляет отчеты по количеству сообщений с упоминаниями ключевых слов, авторов, источников, тональности.

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

Socialbakers (www.socialbakers.com) – объединенная платформа маркетинга на основе анализа социальных сетей, сервис сбора статистики о работе социальных сетей, называющий себя «сердцем статистики Facebook». Система

Socialbakers известна своими рейтингами брендов на Facebook, в разных категориях. Кроме сервис Socialbakers предоставляет возможность мониторинга информации в таких сетях, как в Twitter, Youtube, LinkedIn.

PeerIndex (www.peerindex.net) – бесплатный сервис анализа социальных медиа, прежде всего Twitter. Определяет размеры «социального капитала» или влиятельности компании, профессионала, публикации и др. Предлагает в распоряжение наибольшую базу данных пользователей Twitter, что позволяет обнаруживать сетевые сообщества на основе интересов, демографических данных, местоположения, профессии.

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

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

Babkee (www.babkee.ru) – комплексная автоматизированная система мониторинга социальных медиа. Позволяет решать такие задачи, как защита репутации компании, формирование положительного имиджа, а также выполнять маркетинг и PR, услуги в сфере Social Media Marketing.

SemanticForce (www.semanticforce.net) – сервис, обеспечивающий мониторинг неструктурированных источников – комментариев в сетевых СМИ и интернет-магазинах (Рис. 80). В системе используется технология SemanticForce W3Monitor, которая отслеживает изменения на любых ресурсах, включая сайты без RSS, фрагменты страниц, комментарии к публикациям и дискуссии на форумах. SemanticForce автоматически индексирует тексты статей, на которые ссылаются сообщения (твиты) в микроблогах. Это позволяет находить косвенные упоминания об объекте и значительно расширить по-

крытие. Для мониторинга популярных социальных сетей: Facebook, VKontakte, GooglePlus используются собственные поисковые алгоритмы SemanticForce. Учитываются морфологические особенности и специфика конкретной сети, что позволяет значительно увеличить объем отслеживаемых упоминаний.

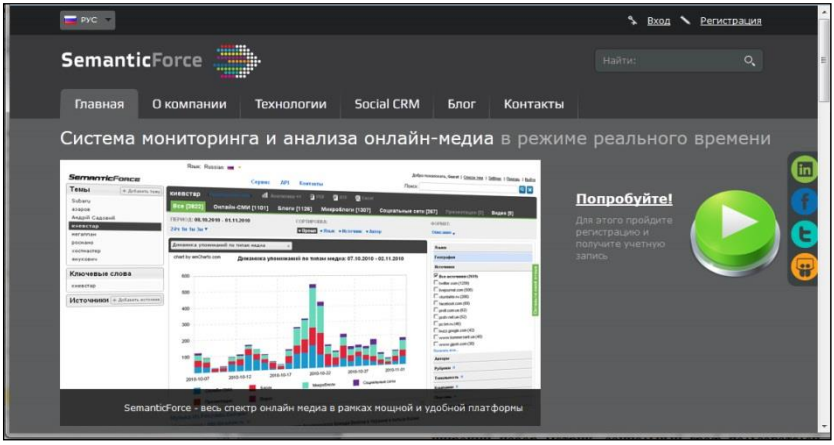


Рис. 80 – Фрагмент веб-сайта службы SemanticForce

Для мониторинга высокочастотных объектов используется технология Twitter Firehose, которая позволяет получать данные из Twitter без задержек по времени и ограничений по объему выгружаемой информации. Технология SemanticForce позволяет отслеживать поведение автора упоминаний и его отношение к объекту мониторинга, автоматически призводить поиск его профилей в Интернет и собирать историю с целью последующего анализа и вовлечения. Также возможна Гео-сегментация, определение географического местоположения автора сообщения. Для анализа сообщений используется детектирование объектов – Автоматическая выборка и статистика по упоминающимся в текстах компаниям, продуктам и персонам. Иерархическая кластеризація обеспечивает навигацию по большому массиву данных, выделяя кластеры по отдельным словам, которые часто упоминаются в контексте с объектами мониторинга. Тональность определяется не для

всего сообщения в системе, а для конкретного объекта в упоминании, что позволяет формировать выборки с различной тональностью - например, в том случае, когда в одном сообщении об определенном бренде говорится позитивно, а о его конкуренте – негативно. В платформе SemanticForce реализована специальная архитектура для хранения, поиска и визуализации комментариев, что позволяет видеть комментарии под исходной статьей или заметкой, к которой они изначально оставлялись. В рамках платформы SemanticForce объединены медиа и веб-аналитика. В систему интегрирован самый популярный сервис веб-аналитики Google Analytics. Аналитические данные из Google Analytics можно найти в отчете по источникам.

Крибрум (www.kribrum.ru) – система мониторинга и анализа социальных сетей, позволяющая отслеживать и анализировать упоминания бренда, продуктов или услуг, ключевых персон, событий, географических названий. Система автоматически определяет эмоциональную окраску высказываний и распределяет публикации по тегам и категориям. Система принадлежит компании «Ашманов и партнёры» и позиционируется как продукт для управление репутацией компании.

3.2.3. Анализ социальных сетей

В качестве примера применения возможностей анализа социальных сетей приведем фрагмент исследования сети связей понятий (фамилий персон), экстрагируемых из корпусов неструктурированных текстов – массивов документов, сканируемых из сети Интернет системой контент-мониторинга InfoStream [Григорьев, 2007].

При построении сети понятий использовались алгоритмы автоматического извлечения понятий из неструктурированных текстов. Следует отметить, что подходы к извлечению различных типов понятий из текстов существенно отличаются как по контексту их представления, так и по структурным признакам. Так, для выявления принадлежности документа к рубрике тематического классификатора могут использоваться специальным образом составленные запросы, включающие логические и контекстные операторы, скобки и т.п.

Выявление географических названий предполагает использование таблиц, в которых помимо шаблонов написания этих названий используются коды стран, названия регионов и населенных пунктов.

Еще один вид понятий, такой как «персоны», экстрагируется из текстов на основании правил, учитывающих таблицы допустимых имен и фамилий, шаблоны инициалов, возможные варианты совместного написания инициалов/имен и фамилий.

Следует отметить, что система InfoStream включает средства извлечения понятий и, среди прочего, предоставляет пользователям результаты в виде «информационных портретов», включающих такие понятия, как ключевые слова, географические названия, фамилии персон, названия фирм и т.д. В рамках этой системы анализируются свойства сетей, образованных понятиями, связанных друг с другом упоминаниями в тех же документах.

Сеть, образованная понятиями, извлекаемыми из потоков текстов, не статичен, а зависит от объемов документов, из которых извлекаются соответствующие понятия. Следовательно, для понимания структуры такой сети необходимо учитывать ее эволюцию.

Ребрам исходной сети приписываются весовые значения, равные количеству документов, в которых встречаются упоминания персон, отвечающих узлам. Для предотвращения «шума», ребра с весом, меньше чем 2 не учитывались. При развитии сети с фиксированным количеством персон, при увеличении количества рассмотренных документов среднее расстояние между узлами, соответственно уменьшается, достигая своего логического насыщения.

Интересен тот факт, что узлы рассматриваемой сети персон с максимальным количеством исходящих ребер преимущественно имеют наибольший уровень посредничества и не могут рассматриваться в качестве основы для построения кластеров при автоматической группировке, а скорее как элементы, соединяющие отдельные группы узлов.

Информационная сеть персон (наиболее упоминаемых) и их связей, полученных путем анализа выходного потока по экономической проблематике за определенный период времени, представлена в виде графа (Рис. 81).

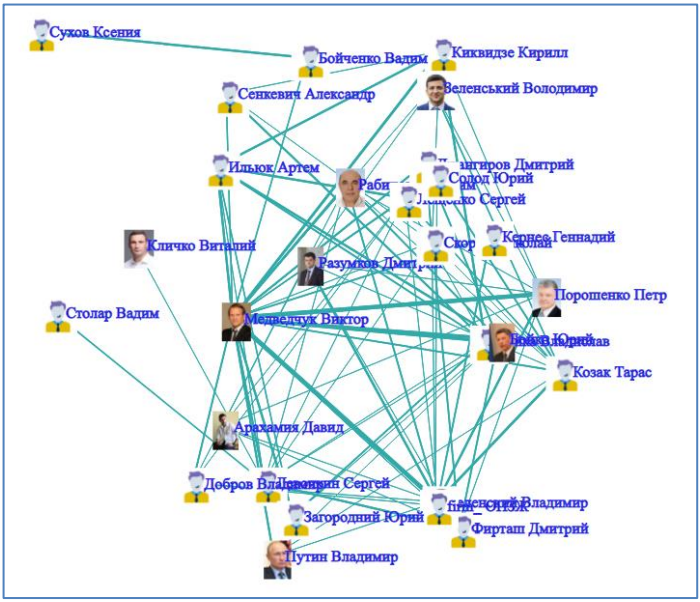


Рис. 81 – Отображение информационной сети персон

Полученные эмпирические результаты могут быть полезными, например, при моделировании экономико-социальных процессов, выявлении и визуализации неявных связей отдельных объектов или субъектов. Феномен стабилизации сети связей на практике позволяет путем анализа небольшого массива документов выявлять устойчивые связи, снижать влияние шумовых факторов. Вместе с тем пока остается открытым вопрос оценки корреляции полученных информационных взаимосвязей персон, рассчитанных путем подсчета частоты документов, в которых персоны упоминаются совместно, и взаимосвязей реальных.

3.3. Глубинный веб, специальные базы данных

Последние исследования веб-пространства показали, что доступные через традиционные информационно-поисковые системы более триллиона веб-страниц – это лишь «поверхностная видимая часть айсберга».

Важной проблемой является поиск информации в «скрытом» или «глубинном» веб-пространстве, где, как было замечено выше, содержится несравнимо большее количество данных, потенциально интересных для конкурентной разведки, чем в открытой части Интернета.

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

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

Как уже было отмечено ранее, необходимой (в том числе и для конкурентной разведки) информации в сети Интернет значительно больше, чем ее охватывают универсальные поисковые машины.

Предполагается, что в отличие от «познаваемой» части сети Интернет, «скрытая» часть оказалась в сотни раз более объемной.

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

3.3.1. Понятие «глубинный веб»

Совокупность источников в веб-пространстве, недоступных пользователям традиционных поисковых систем, образует так называемый «глубинный веб» – понятие, введенное Джиллом Иллсвортом (Jill Ellsworth) в 1994 г. Т.е. под глубинным веб (invisible web, deep web, hidden web) принято понимать ту часть веб-пространства, которая не индексируется роботами (web crawlers) поисковых систем. Используя аналогию, информация, будучи недоступной для поиска, находится

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

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

Сегодня ситуация изменилась, например, ведущие поисковые системы могут индексировать документы, представленные в форматах, содержащих текст. Конечно, это, прежде всего, pdf, rtf и doc. В 2006 году Google запатентовала способ поиска в глубинном веб: «Searching through content which is accessible through web-based forms» (Рис. 82). По мнению разных авторов к видимому веб относится лишь 20–30 % веб-пространства.

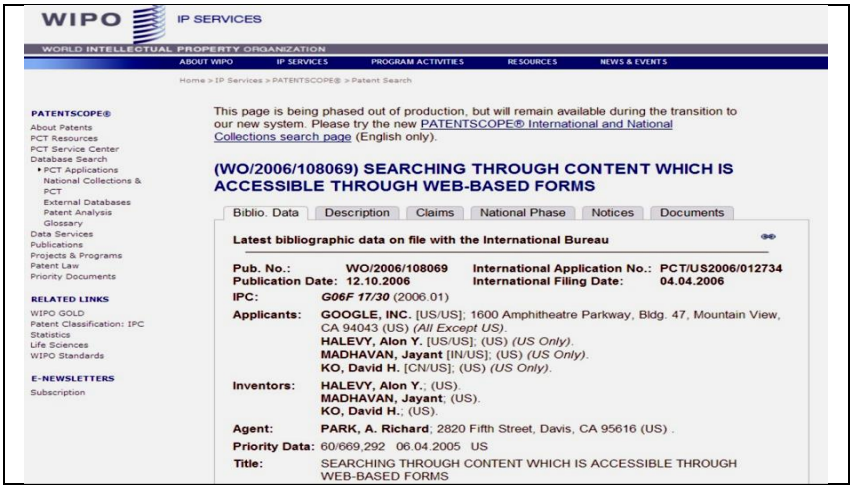


Рис. 82 – Фрагмент веб-ресурса WIPO с описанием патента Google на поиск в глубинном веб

3.3.2. Причины возникновения

В глубинном веб находятся веб-ресурсы, не связанные с остальными ресурсами гиперссылками – например, страницы, динамически создаваемые по запросам к базам данных, документы из баз данных, доступные пользователям через поисковые веб-формы (но не по гиперссылкам). Такие документы остаются недоступными для робота, неспособного в режиме реального времени правильно заполнить поля формы значениями (формировать запросы к базам данных).

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

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

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

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

Общепризнано, что ценность ресурсов глубинного веб иногда выше ценности ресурсов видимой части веб-пространства.

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

Многие компании сначала подключаются к общей Сети, и лишь потом тратят большие средства на защиту. Владельцы сайтов могут попытаться запретить индексацию тех или иных страниц своих ресурсов, прописав запрещающую команду в файле robots.txt, но поисковые системы могут ее проигнорировать. Поэтому такие ресурсы либо удаляют, либо удаляют гиперссылки, переводя ресурсы в глубинный веб. Например, некоторые бизнес-каталоги отказываются отдавать свои объявления роботам поисковых систем, т.е. защищая свои информационные активы компании переводят свои ресурсы в глубинный веб.

3.3.2. Виды ресурсов глубинного веб

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

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

К глубинному веб также относятся многочисленные системы интерактивного взаимодействия с пользователями – помощи, консультирования, обучения, требующие участия людей для формирования динамических ответов от серверов. К ним также можно отнести и закрытую (полностью или частично) информацию, доступную, пользователям Сети только с определенных адресов, групп адресов, иногда городов или стран. К «скрытой» части Сети многие причисляют и веб-страницы, зарегистрированные на бесплатных серверах, которые индексируются, в лучшем случае, лишь частично – поисковые системы во избежание рекламного спама не стремятся обходить их в полном объеме.

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

3.3.2. Ресурсы глубинного веб

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

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

Для поиска в глубинном веб, а именно в том его сегменте, который составляют базы данных, сегодня уже существуют некоторые специализированные ресурсы. Лидером среди навигаторов в глубинном веб является сайт CompletePlanet (www.completeplanet.com) компании BrightPlanet. Этот сайт является крупнейшим каталогом, насчитывающим свыше 100 тысяч ссылок. Компания BrightPlanet также создала персональную утилиту для поиска в онлайновых базах данных LexiBot, которая может обеспечивать поиск в нескольких тысячах поисковых систем «глубинного» веб. Метапоисковый пакет DeepQueryManager (DQM) этой же компании обеспечивает поиск более чем по 70 тысячам «скрытым» веб-ресурсам.

Исследование, проведенное еще в 2006 г. [Не, 2007] показало, что глубинный веб охватывает более 300 тыс. сайтов, связанных с более 450 тыс. базами данных, не охватываемых традиционными поисковыми системами. К наиболее интересным для бизнес-аналитиков ресурсам глубинного веб относятся: базы данных юридических и физических лиц; отрасле-

вые базы данных; репутационные базы данных (черные и белые списки); криминологические базы данных; базы данных товаров и услуг; каталоги продукции и т.п. К всемирно известным бизнес-ресурсам, размещенным в глубинном веб, относятся: amazon.com, ebay.com, realtor.com, cars.com, imdb.com.

Приведем еще несколько примеров баз данных и каталогов глубинного веб:

FindLaw (www.findlaw.com) – один из наиболее популярных в мире юридических веб-сайтов – большой каталог правовых ресурсов, содержащий аннотированный список свободно доступных баз данных нормативно-правовых документов, для которых данный ресурс является «точкой входа». Фрагмент веб-сайта сервиса FindLaw приведен на Рис. 83.

Рис. 83 – Фрагмент веб-сайта сервиса FindLaw

Politicalinformation.com (www.politicalinformation.com) – ресурс для журналистов, политиков, студентов и политических деятелей, сервис, обеспечивающий оперативный поиск в

5000 отобранных веб-сайтах политической направленности, предоставление новостей из нескольких десятков авторитетных источников.

Академическая поисковая система Biefield Билефельда BASE (https://archive-it.org/) – одна из крупнейших в мире поисковых систем академических веб-ресурсов. Обеспечивается доступ к полным текстам около 60% проиндексированных документов бесплатно (открытый доступ). BASE находится в ведении Библиотеки Университета Билефельда.

CiteSeerX (https://citeseerx.ist.psu.edu/index) - это электронная библиотека научной литературы и поисковая система. Сервис создан для распространения научной литературы и улучшения функциональности, удобства использования, доступности, стоимости, полноты, эффективности и своевременности доступа к научным и научным знаниям.

Data.gov (https://www.data.gov/) – «дом данных». Согласно условиям Федеральной политики открытых данных 2013 г., вновь созданные правительственные данные должны быть доступны в открытых машиночитаемых форматах, при этом сохраняется конфиденциальность и безопасность.

Mednar (https://mednar.com/mednar/desktop/en/search.html) – бесплатная поисковая система в глубокой сети, ориентированная на медицину. Поскольку Mednar является общедоступным поиском, он не может получить результаты из личной подписки или дополнительных медицинских ресурсов.

World Wide Sicence (https://worldwidescience.org/) - глобальный научный портал, состоящий из научных баз данных и порталов.

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

Традиционная поисковая система чаще всего может выдать адрес базы данных, но не скажет, какие документы конкретно содержаться в ней. Типичный пример – информаци-

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

Парадоксально, но в качестве одного из ресурсов глубинного веб можно рассматривать и архив ресурсов открытого веб-пространства. Такой архив – Internet Archive создается с 1996 (www.archive.org). Сегодня объем базы данных Alexa превышает 538 млрд. веб-страниц (Рис. 84), 28 млн. книг и текстов, 14 млн. аудиозаписей, 6 млн. видео, 3,5 млн. изображений, 580 тысяч компьютерных программ.

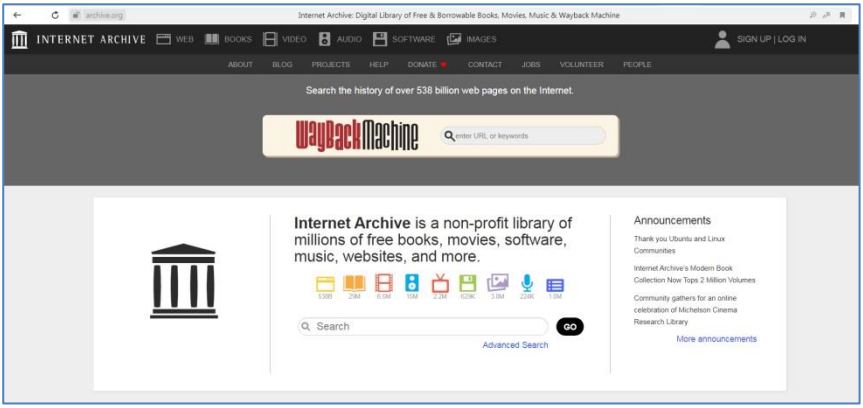


Рис. 84 – Заглавная страница веб-сайта www.archive.org

Технология хранилища Internet Archive включает ряд современных средств управления гигантским документальным хранилищем. Например, с помощью этой технологии выполняется кластеризация веб-ресурсов, т.е. формирование коллекций документов, близких по тематикам. Особый интерес у пользователей сервиса Internet Archive вызывает «Машина времени» (Wayback Machine), открывающая доступ к временным срезам веб-пространства. Одно из наиболее интересных практических применений этой технологии – восстановление документов, некогда опубликованных в веб-пространстве, но впоследствии удаленных. При этом рост глубинного веб грозит серьезными проблемами полноты в хранилище системы,

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

3.3.3. Сервисы работы с глубинным веб

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

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

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

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

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

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

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

3.3.4. Специальные базы данных

Как правило, для успешного ведения конкурентной разведки должен быть создан и поддерживаться банк данных, включающий следующие основные базы данных [Ландэ, 2005]:

1. Конкуренты (действующие и потенциальные);
2. Информация о рынке (тенденции, номенклатурная, ценовая, адресная информация);
3. Технологии (продукты, выставки, конференции, ГОСТы, качество);
4. Ресурсы (сырье, человеческие и информационные ресурсы);

5. Законодательство (международные, центральные, региональные и ведомственные нормативно-правовые акты);

6. Общие тенденции (политика, экономика, региональные особенности, социология, демография).

Сегодня для конкурентной разведки основными источниками информации служат Интернет, пресса, а также открытые базы данных. Но если доступ к обычным интернет-ресурсам можно считать условно бесплатным, то, в большинстве случаев, доступ к базам данным требует не только регистрации, но и оплаты таких услуг. Кроме того практически все они могут быть отнесены к так называемому «скрытому» веб-пространству.

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

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

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

Крупнейшая в мире полнотекстовая онлайновая информационная система ***LexisNexis*** (www.lexisnexis.com), которая содержит свыше 2 миллиардов документов из 45 тыс. источников с архивом глубиной более 30 лет по бизнес-информации и более 200 лет по правовой информации, относится к разряду «скрытого» веб (Рис. 85). Каждую неделю в архивы добавляется еще 14 млн. документов. В отличие от неструктурированных массивов «поверхностного» веб, поль-

зователи LexisNexis могут использовать мощные инструменты поиска для получения достоверной и классифицированной информации.

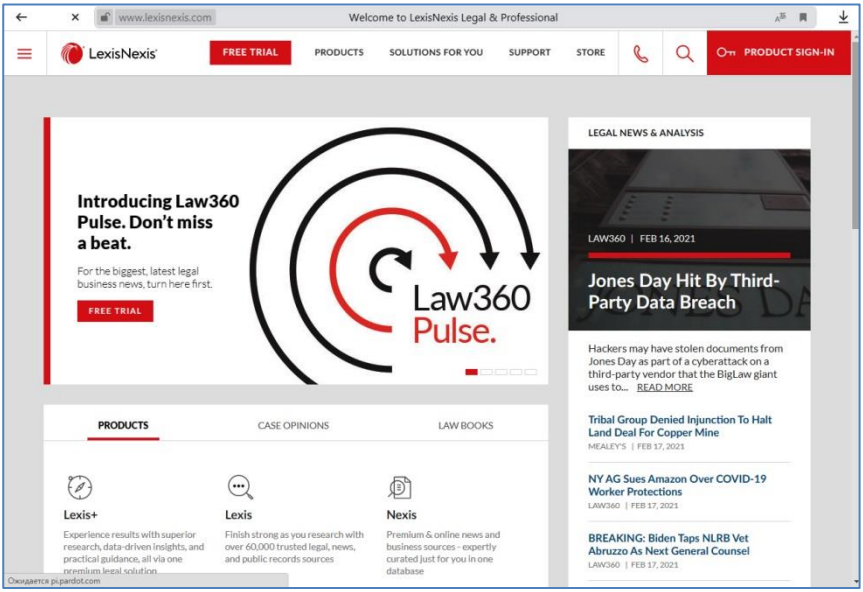


Рис. 85 – Фрагмент веб-сайта службы LexisNexis

Служба ***Factiva*** (global.factiva.com), подразделение компании Dow Jones, занимается предоставлением доступа к деловой и аналитической информации. В основе службы Factiva имеется более 35 тыс. первичных источников из 159 стран мира. В базе данных Factiva содержатся материалы более чем по 36,5 млн. компаний, а также полная подборка информации Investext.

Компания ***Internet Securities*** (www.internetsec.com), бренд ISI Emerging Markets, охватывает 80 тематических информационных разделов, формируемых из 16 тыс. источников информации – тексты статей, финансовые и аналитические отчеты, корпоративная информация, макроэкономическая статистика, данные по рынкам (Рис. 86). Основные продукты ISI Emerging Markets: CEIC Data, Emerging Market In-

formation Service (EMIS), Islamic Finance Information Service (IFIS), IntelliNews, ISI Compliance Edition, ISI DealWatch.

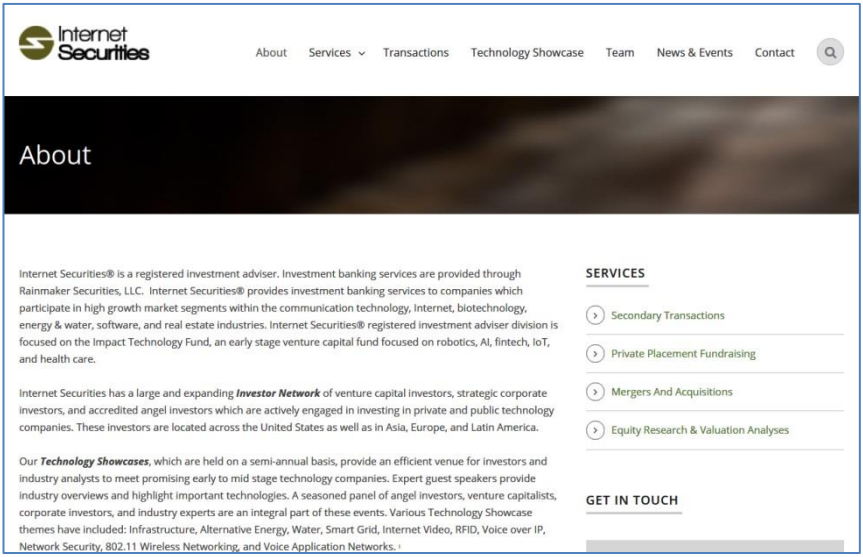


Рис. 86 – Фрагмент веб-сайта службы Internet Securities

Europages (www.europages.eu) – Европейская бизнес-директория – информационно-поисковая B2B-система, охватывающая свыше 3 млн. поставщиков, производителей и дистрибьюторов в Европе и во всем мире.

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

Наряду с базами данных, одним из самых эффективных источников информации могут служить отчеты и справки аутсорсинговых компаний, профессионально занимающихся конкурентной разведкой и сбором сведений о коммерческих

структурах и рынках. Их продукция, на самом деле, и является результатом конкурентной разведки.

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

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

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

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

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

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

4. Репутационный анализ

4.1. Проблема управления репутацией

Репутация представляет собой социальную оценку группы субъектов о человеке, группе людей или компании, сформировавшуюся на основе некоторых критериев.

Репутация компании – это комплекс оценочных представлений целевой аудитории о компании, сформированный на основе факторов репутации, имеющих значение для этой аудитории. Согласно информационному письму Высшего хозяйственного суда Украины «О некоторых вопросах практики применения хозяйственными судами законодательства об информации» от 28 марта 2007 деловую репутацию юридического лица составляет престиж ее фирменного (коммерческого) наименования, торговых марок и других принадлежащих ему нематериальных активов среди круга потребителей ее товаров и услуг.

Успех компании напрямую связан с ее репутацией. Так исследование, проведенное австралийскими учеными П. Робертсом и Г. Даулинг, выявило, что чем выше репутация у компании, тем, во-первых, дольше период, в течение которого она получает максимальный доход от своей деятельности, и, во-вторых, тем меньше времени кого компании нужно для достижения средних по отрасли финансовых показателей при внедрении инноваций. Репутационный капитал (Reputational Capital) – понятие не только маркетинговое, не меньшее отношение оно имеет и к финансам. Денежный эквивалент деловой репутации может быть выражен в форме гудвилла (goodwill). Соответствии с Международными стандартами финансовой отчетности (МСФО) гудвил, представляет собой разницу между ценой, заплаченной за предприятие покупателями, и «справедливой стоимости» (данная величина часто значительно отличается от простой стоимости всех активов компании). Например, в правилах бухгалтерского учета под репутацией понимается «разница между покупной ценой организации и стоимостью по балансу всех ее активов и обязательств».

Финансовая отдача компании напрямую связаны с ее репутацией. Так исследование, проведенное австралийскими учеными Г. Даулингом и П. Робертсом [Roberts, 2002], выявило два преимущества благоприятного имиджа компании. Сравнив данные рейтинга 500 лучших и наиболее уважаемых компаний США, ежегодно составляемого американским журналом Fortune, за 1984–1995 годы с финансовыми показателями компаний за этот же период, ученые выявили взаимосвязь между репутацией фирмы и ее финансовым уровнем. Выяснилось, что чем выше репутация у компании, тем, во-первых, дольше период, на протяжении которого она получает максимальный доход от своей деятельности, и, во-вторых, тем меньше времени компании нужно для достижения средних по отрасли финансовых показателей при внедрении инноваций.

Чтобы иметь возможность выяснять нематериальную цену компании, разрабатываются экспертные оценки репутации. Стоимость репутации может определяться экспертами, например, таким образом. Сначала рассчитывается доход, полученный компанией за счет бренда (разница между реальной прибылью и доходами, которые можно получить, продавая небрендированный товар), а затем полученная сумма умножается на специально рассчитанный коэффициент (зависящий от положения компании в отрасли, стабильности финансовых показателей и т.д.) Результат и есть цена брэнда, являющегося важной частью репутации.

Существуют и косвенные оценки уровня репутации компаний, например, основанные на результатах опроса руководителей фирм и аналитиков, оценивающих компании по таким параметрам, как качество менеджмента и продукта, способность привлечь и удержать квалифицированные кадры, финансовая стабильность, эффективное использование активов, инвестиционная привлекательность, применение новых технологий и т.п.

Понятие «Управление репутацией в Интернете» (Online Reputation Management, ORM) по сути представляет собой комплекс мероприятий по обнаружению в сети негативного контента и сведения его к минимуму в социальных медиа и в результате поисковой выдачи. Это, своего рода, PR-кампания в киберпространстве. Ветвью ORM является SERM (Search

Engine Reputation Management) – управление репутацией в поисковых системах. На Западе такие услуги практикуются очень активно, и рост ORM в год составляет порядка 35–40 %.

Сегодня по статистике компании Google 70 % пользователей ищут отзывы о товарах и услугах, прежде чем купить их. Историически первой компанией, которая стала практиковать двустороннюю связь с клиентами в социальных сетях, стала компания eBay. На основе обратной связи был составлен рейтинг продавцов, на который могли опираться покупатели при принятии решения о покупке. В России ярким примером отображения репутации компании, базирующейся на отзывах пользователей, можно назвать систему Яндекс.Маркет. Больше половины пользователей Интернета при выборе того или иного продукта, компании, заказчика, исполнителя и т.д. опираются на информацию, предоставленную другими пользователями.

Работы по управлению репутацией проводят как специализированные PR-агентства, работающие на просторах вебпространства, так и подразделения SEO-агентств, которые запускают PR-кампании, направленные на поиск и устранение негативного контента. Кроме того, такие услуги предоставляют и частные лица – фрилансеры, специалисты в области интернет-маркетинга и SEO. В крупных компаниях существуют свои собственные отделы, работа которых направлена на управление репутацией фирмы, бренда, товара, услуги.

Понятие «Управление репутацией в Интернете» (ORM) уже стало устоявшимся термином и на эти цели на Западе ежегодно выделяется часть бюджета большинства крупных компаний. Вместе с ростом влияния социальных медиа на взгляды и предпочтения людей растет и необходимость крупных компаний следить за своим имиджем в сети. На этом фоне не кажется странным рост рынка ORM на 40 % ежегодно.

Основная задача управления репутацией – формирование положительного имиджа о компании и ее продукте. Так как сложно охватить абсолютно все пользовательские отзывы и убрать весь негатив, обычно усилия концентрируются в трех областях: поисковой выдаче, отзывах в электронных СМИ и упоминаниях в социальных медиа. Приходится работать как с контентом, создаваемым редакторами различных изданий, так и простыми пользователями. Для создания це-

лостного положительного образа информация из этих трех источников должна быть положительной или нейтральной.

Управление репутацией в поисковых системах – Search Engine Reputation Management (SERM) – комплекс мероприятий, направленных на исключение негативных отзывов о компании, товаре или услуге в результатах выдачи поисковой системы.

Услуга управления репутацией в поисковых системах необходима:

- компаниям, желающим исключить или минимизировать негативные отзывы о своей деятельности (продукции);
- компаниям, желающим сформировать положительные отзывы или увеличить их количество и видимость для целевой аудитории.

Негативная информация, наносящая вред репутации в сети, может быть различного происхождения. Условно выделяют три основные группы происхождения негативного контента [Ковальчук, 2012]:

- неумышленный негатив – это могут быть как отзывы недовольных клиентов, которые не имеют помыслов нанести вред репутации компании, а просто не удовлетворены итогами сотрудничества, так и неосторожно размещенные в Интернете фотографии с корпоративных праздников, высказывания сотрудников в адрес клиентов и т. п. Обычно такой негатив не представляет большой угрозы, но игнорировать его ни в коем случае нельзя;
- умышленный негатив с целью ударить по репутации – в этом случае классический пример – отрицательные отзывы уволенных или уволившихся сотрудников, недовольных концепцией компании.
- черная PR-кампания – самый опасный вид негативного контента, наносящий серьезный удар по репутации. Такие PR-кампании проводят специалисты, которые тщательно изучают бизнес конкурента и точно знают, где скрыта ахиллесова пята. Организуются крупные рейдерские захваты, способные привести к полному краху не только репутацию, но и весь

бизнес в целом. Данную услугу у PR-специалистов заказывают крупные серьезные конкуренты.

Самыми уязвимыми тематиками в плане притяжения негативных отзывов можно назвать:

- банки, финансовые институты;
- деятели политики и шоу-бизнеса;
- туризм, путешествия (отзывы об отелях, курортах, туроператорах, авиаперевозчиках);
- мобильная техника и связь (операторы, телефоны, электронные планшеты);
- бытовая техника;
- заведения общепита (рестораны, кафе, бары).

Как пространство мониторинга для управления репутацией выбирают сетевые ресурсы, где размещаются отзывы потребителей:

- социальные сети, мессенджеры;
- блоги и форумы;
- тематические веб-сайты и порталы;
- специальные сервисы отзывов.

Продвигаются страницы с позитивным контентом при помощи стандартных инструментов поисковой оптимизации (Search Engine Optimization, SEO), таких как ссылочные биржи, покупка, обмен ссылками на статьи с тематическими ресурсами, размещение анонсов, новостей и др. При этом позитивный контент следует размещать регулярно, так как негативный контент способен проявляться вновь и портить репутацию.

Бороться с негативным контентом призвана поисковое управление репутацией - SERM. Задача SERM состоит в вытеснении из результатов поиска веб-страниц с нежелательной информации, в результате чего целевая аудитория перестанет видеть такие ресурсы, так как пользователи не будут выходить на них с помощью поисковых систем. Для достижения этой цели создаются материалы с положительным контентом, предполагая, что они вытеснят негативные нежелательные сообщения.

Управление репутацией в сети обычно начинают с мониторинга поисковой выдачи и социальных медиа с целью выявления информации по заданному объекту. Существует несколько методов мониторинга:

- ручной мониторинг поисковых систем путем ввода целевых поисковых запросов;
- использование систем оповещения, интегрированных с поисковыми системами, например, Яндекс.Блоги (blogs.yandex.ru) и Google Оповещения (google.com/alerts). В этих случаях релевантная информация поступает на электронную почту подписчика;
- использование специальных средств мониторинга репутации компаний в социальных сетях.

В качестве пространства мониторинга для управления репутацией выбирают сетевые ресурсы, где размещаются отзывы потребителей:

- социальные сети;
- блоги и форумы;
- тематические веб-сайты и порталы;
- специальные сервисы отзывов.

Одним из критериев качества услуги мониторинга репутации является полнота охвата – доля информации об объекте, исследуемая во время работы от общего объема информации в сети об объекте. По-прежнему основным инструментом поиска информации являются традиционные поисковые системы, они охватывают значительную часть интернет-контента, а также некоторую часть социальных медиа.

Сегодня во всем мире существуют сотни систем мониторинга репутации, среди которых можно назвать системы Babkee, Brand-spotter, BuzzLook, Buzzware, IQBuzz, Крибрум, SemanticForce, Wobot, Youscan. В исследованиях Кена Барбери (Ken Burbary) и Адама Коэна (Adam Cohen) [Burbary, 2009-2013] приведен список из 230 систем мониторинга репутации, для многих из которых предлагаются бесплатные тестовые периоды для оценки качества их работы.

4.2. Моделирование репутации в сетях

В последнее время в рамках теории анализа социальных сетей большое внимание уделяется оценке репутации отдельных субъектов (агентов, узлов социальных сетей) и уровня доверия между ними [Расторгуев, 2006], [Губанов, 2009].

Формально социальная сеть представляет собой граф, в котором множество вершин – это совокупность агентов, субъектов – индивидуальных или коллективных, а множество ребер представляет собой совокупность отношений, совокупности социальных связей между агентами.

При моделировании социальных сетей возникает необходимость учета динамики социальных связей – взаимного влияния агентов.

Влияние в данном случае рассматривается как процесс и результат изменения индивидом (субъектом влияния) поведения другого субъекта – объекта влияния, его установок и оценок в ходе взаимодействия [Расторгуев, 2006]. Таким образом, влияние – это способность воздействовать на чьи-либо представления или действия [Новиков, 2002]. Различают направленное и ненаправленное влияние [Новиков, 2007]. Направленное влияние использует в качестве механизмов воздействия на другого человека убеждение и внушение. При этом индивид – субъект влияния – ставит перед собой задачу добиться определенных результатов от объекта влияния. Ненаправленное (нецеленаправленное, косвенное) влияние – это влияние, при котором индивид не ставит перед собой задачу добиться определенных результатов от объекта влияния.

Целенаправленное влияние участников социальной сети (или субъектов, не входящих в сеть, но использующих ее в качестве инструмента информационного воздействия) является частным случаем информационного управления, заключающегося в формировании у управляемых субъектов такой информированности, чтобы принимаемые ими на ее основе решения были наиболее выгодны для управляющего субъекта.

Возможности влияния одних участников социальной сети на других ее участников существенно зависят от репутации первых. Репутация – «создавшееся общее мнение о достоинствах или недостатках кого-либо, чего-либо, общественная оценка» [Ермаков, 2005]. Репутацию можно рассматривать как «весомость» мнения сообщества об отдельном агенте или группе агентов, определяемую его взглядами и деятельностью

(активностью). При этом репутация может быть как индивидуальной, так и коллективной.

Репутация возрастает, если выбор агента (ответы на некоторые ключевые вопросы) совпадает с тем, чего от него ожидает сообщество, и понижается в противном случае.

Приведем формальное определение репутационной модели [Губанов, 2009].

Пусть $\{a_1, a_2, ..., a_n\}$ – множество агентов – узлов социальной сети, которые влияют друг на друга. Матрицу влияния обозначим как $A = \|a_{ij}\|_{i=1,n}^{j=1,n}$ ($a_{ij} \geq 0$ обозначает степень доверия *i*-го агента *j*-му). При этом очевидно, что если *i*-й агент влияет на *j*-го, а *j*-й влияет на *k*-го, то это означает следующее: *i*-й агент косвенно влияет на *k*-го (транзитивность), что позволяет строить цепочки косвенных влияний.

Предположим, что у каждого агента в начальный момент времени имеется мнение по некоторому ключевому вопросу. Пусть мнение сообщества агентов сети отражает вектор начальных мнений b^0 размерности *n*. Мнение каждого агента меняется под влиянием мнений других агентов социальной сети.

Будем считать, что мнение *i*-го агента в момент времени t равно

$$b_i^t = \sum_{j=1}^{n} a_{ji} b_j^{t-1}$$

В [Ермаков, 2005] показано, что при многократном обмене мнениями, мнения агентов сходятся к результирующему вектору мнений $B = \lim_{t \to \infty} b^t$. Таким образом, справедливо соотношение $B = Ab$.

Обозначим r_i – параметр, описывающий репутацию *i*-го агента в социальной сети (сообществу), которую можно определить как нормированную сумму его влияний на всех остальных агентов социальной сети (предполагается, $a_{ij} \geq 0, \ \ i, j = 1, ..., n$), т.е.

$$r_i = \frac{\sum_{i \neq j} a_{ij}}{R}, \quad j = 1,...,n.$$

Здесь $R = \sum_k \sum_{j \neq k} a_{kj}$, $k, j = 1,...,n$ – суммарное взаимное влияние друг на друга всех членов социальной сети.

В соответствии с приведенным выражением агент i имеет тем более высокую репутацию, чем выше его влияние на всех остальных членов социальной сети.

Моделирование с использованием гиперкомплексных числовых систем (ГЧС) позволяет применять развитый инструментарий из этой области математики.

В рамках модели, основанной на использовании ГЧС, каждый субъект (узел социальной сети) характеризуется своим отношением к ряду важных (ключевых) вопросов (пусть их количество равно N). Тогда, по аналогии с [Lande, 2012], субъекту A можно поставить в соответствие гиперкомплексное число с базисом размерности $2N$:

$$A = e_1 w_1^+ + e_2 w_1^- + ... + e_{2N-1} w_N^+ + e_{2N} w_N^- .$$

При этом каждому вопросу приписываются весовые значения w_i^+ и w_i^-, которые соответствуют уровню положительного отношения субъекта к данному вопросу (w_i^+) или отрицательного (w_i^-), что является естественным расширением приведенного выше подхода. Оба значения могут быть в интервале [0, 1], в отдельных случаях можно предположить, что $w_i^+ + w_i^- = 1$.

Предлагается использовать ГЧС размерности $2N$ с базисом $\{e_1, e_2, ..., e_{2N}\}$ и законом умножения, который можно представить в виде таблицы:

	e_1	e_2	e_3	e_4	⋯	e_{2N-1}	e_{2N}
e_1	e_1	e_2	0	0	…	0	0
e_2	e_2	e_1	0	0	…	0	0

e_3	0	0	e_3	e_4	…	0	0
e_4	0	0	e_4	e_3	…	0	0
…	…	…	…	…	…	…	…
e_{2N-1}	0	0	0	0	0	e_{2N-1}	e_{2N}
e_{2N}	0	0	0	0	0	e_{2N}	e_{2N-1}

Модель субъекта социальной сети в данном случае рассматривается как гиперкомплексное число вида:

$$D = e_1 w_1^+ + e_2 w_1^- + e_3 w_2^+ + e_4 w_2^- + ... + e_{2N-1} w_N^+ + e_{2N} w_N^- .$$

Можно рассмотреть оценку близости мнений двух субъектов $Est(A,B)$ $A = e_1 a_1^+ + e_2 a_1^- + ... + e_{2N-1} a_N^+ + e_{2N} a_N^-$ и $B = e_1 b_1^+ + e_2 b_1^- + ... + e_{2N-1} b_N^+ + \ + e_{2N} b_N^-$:

$$Est(A,B) = Norm(\frac{1}{N}(\sum_{i=1}^{N}(e_{2i-1}a_i^+ + e_{2i}a_i^-)(e_{2i-1}b_i^+ + e_{2i}b_i^-)) ,$$

где $Norm(\bullet)$ – функция нормы гиперкомплексного числа: $Norm(e_{2i-1}) = e_1,\ \ Norm(e_{2i}) = -e_1$.

Отношение большей части участников социальной сети (общества) к выбранным вопросам также представляется в виде гиперкомплексного числа $Q = q_1 e_1 + q_2 e_2 + q_3 e_3 + ... + q_{2N} e_{2N}$, как и отдельный субъект D. Чем больше значение $Est(Q,D)$, тем субъект более лояльный, «релевантный» обществу.

Можно ввести и другую оценку близости между гиперкомплексными числами, по аналогии с нормой различий между обычными векторами в векторном пространстве:

$$Est_1(A,B) = Norm(\frac{1}{N}(\sum_{i=1}^{N}(e_{2i-1}a_i^+ - e_{2i-1}b_i^+)^2(e_{2i}a_i^- - e_{2i}b_i^-)^2) .$$

В этом случае субъект будет более лояльным по отношению к обществу, если оценка $Est_1(Q,D)$ будет меньшей.

Вместе с этим, вторая оценка по содержанию менее пригодна для задач выявления взаимного влияния субъектов.

Например, при сравнении отношения субъекта к обществу с запросом, даже при полностью нулевых значениях весовых коэффициентов, относящихся к значению всей социальной сети (обществу), сумма в приведенном для выражения $Est_1(Q,D)$ не будет нулевой, т.е. полностью зависит от субъекта. Поэтому ограничимся применением первой оценки.

Рассмотрим для примера некоторые упрощенные частные случаи, при которых анализируются отношения общества и субъекта к одному вопросу.

1. Пусть значение, соответствующее обществу, имеет вид: $Q=\frac{1}{2}e_1+\frac{1}{2}e_2$, т.е. отношение к выбранному вопросу в обществе может быть как позитивным, так и негативным, с равной вероятностью. Пусть отношение субъекта к этому же вопросу однозначно позитивное, а именно: $D=e_1$. В этом случае $Est(Q,D)=0$, что соответствует полной неопределенности.

2. Пусть значение, соответствующее обществу, как и в предыдущем случае, имеет вид: $Q=\frac{1}{2}e_1+\frac{1}{2}e_2$. Пусть отношение субъекта к этому же вопросу также имеет вид: $D=\frac{1}{2}e_1+\frac{1}{2}e_2$. В этом случае $Est(Q,D)=\frac{1}{4}+\frac{1}{4}-\frac{1}{2}=0$, что, как и в предыдущем случае, соответствует полной неопределенности.

3. Пусть значение, соответствующее обществу, имеет вид: $Q=e_1$, а отношение субъекта к этому же вопросу: $D=\frac{4}{5}e_1+\frac{1}{5}e_2$, тогда $Est(Q,D)=\frac{4}{5}-\frac{1}{5}=\frac{3}{5}$.

Следует отметить, что не все нулевые элементы приведенной выше «идеальной» таблицы в реальности могут быть нулевыми, однако предполагается, что данная таблица будет разреженной. Редкие ненулевые элементы в ней могут характеризовать взаимосвязь различных вопросов.

Значения коэффициентов при базисных элементах образов субъектов социальной сети могут соответствовать вероятностям позитивного (или негативного) отношения субъектов к соответствующим вопросам. В этом случае путем перенуме-

рации базисов таблицу умножения ГЧС можно представить в следующем виде:

<table>
<tr><td></td><td>e_1</td><td>e_2</td><td>e_3</td><td>e_4</td><td>…</td><td>e_{4N-3}</td><td>e_{4N-2}</td><td>e_{4N-1}</td><td>e_{4N}</td></tr>
<tr><td>e_1</td><td colspan="4" rowspan="4">B_1</td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>e_2</td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>e_3</td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>e_4</td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>…</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>…</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>e_{4N-3}</td><td></td><td></td><td></td><td></td><td></td><td colspan="4" rowspan="4">B_N</td></tr>
<tr><td>e_{4N-2}</td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>e_{4N-1}</td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>e_{4N}</td><td></td><td></td><td></td><td></td><td></td></tr>
</table>

где блок B_1 (а в общем случае и любой ненулевой блок) будет иметь вид, дополненный коэффициентами w_i^j, вычисляемыми в процессе обучения модели:

	e_1	e_2	e_3	e_4
e_1	$w_1^1 e_1$	$w_1^2 e_2$	$w_2^1 e_1$	$w_2^2 e_1$
e_2	$w_1^2 e_2$	$w_1^1 e_1$	$w_2^2 e_1$	$w_2^1 e_1$
e_3	$w_2^1 e_1$	$w_2^2 e_1$	$w_3^1 e_3$	$w_3^2 e_4$
e_4	$w_2^2 e_1$	$w_2^1 e_1$	$w_3^2 e_4$	$w_3^1 e_3$

При этом w_1^1 – вес позитивного отношения к вопросу t_1, w_1^2 – вес негативного отношения к вопросу t_1, w_2^1, w_2^2 – веса

взаимосвязей наличия противоречивых одновременных позитивных и негативных отношений к вопросу t_1.

Тогда для документа $A = a_1e_1 + a_2e_2 + a_3e_3 + a_4e_4$ и документа $B = b_1e_1 + b_2e_2 + b_3e_3 + b_4e_4$ оценка близости будет иметь следующий вид:

$$Sim(A \cdot B) = Norm(\frac{1}{2}(e_1(w_1^1(a_1b_1 + a_2b_2) + w_2^1(a_1b_3 + a_2b_4 + a_3b_1 + a_4b_2) +$$
$$+w_2^2(a_1b_4 + a_2b_3 + a_3b_2 + a_4b_1)) + e_2(w_1^2(a_1b_2 + a_2b_1)) +$$
$$+e_3(w_3^1(a_3b_3 + a_4b_3)) + e_4(w_3^2(a_3b_4 + a_4b_4)))).$$

При этом функция $Norm$ соответствует той ГЧС, таблица которой используется для поиска. Следует учитывать, что предложенная таблица может быть составной частью таблицы большей размерности. Переход от заполненной таблицы умножения к слабозаполненной (разреженной) можно осуществить изоморфным переходом [Калиновский, 2012], что значительно сократит количество операций при вычислении функции близости между субъектами.

В общем случае уровень доверия (близость) между субъектами $A = e_1a_1^+ + e_2a_1^- + ... + e_{2N-1}a_N^+ + e_{2N}a_N^-$ и $B = e_1b_1^+ + e_2b_1^- + ... + e_{2N-1}b_N^+ + e_{2N}b_N^-$, которую можно трактовать как степень доверия, также характеризуется функцией, приведенной выше с учетом возможного наличия ненулевых недиагональных элементов таблицы:

$$Sim(A,B) = Norm\left(\begin{array}{l}\frac{1}{N}\left(e_1a_1^+ + e_2a_1^- + ... + e_{2N-1}a_N^+ + e_{2N}a_N^-\right) \\ +\left(e_1b_1^+ + e_2b_1^- + ... + e_{2N-1}b_N^+ + e_{2N}b_N^-\right)\end{array}\right).$$

Репутация субъекта $A_i = e_1a_{i1}^+ + e_2a_{i1}^- + ... + e_{2N-1}a_{iN}^+ + e_{2N}a_{iN}^-$ в рамках всей социальной сети (т.е. по отношению к обществу) при этом определяется, как нормированная сумма уровней доверия со всеми остальными субъектами:

$$Trust(A_i) = \frac{\sum_{j \neq i} Sim(A_i, A_j)}{\sum_{k \neq j} Sim(A_k, A_j)}.$$

Для оценки уровня взаимного влияния субъектов социальной сети могут также использоваться другие методы, среди которых можно выделить: расчет меры взаимной информации (mutual information), расчет модифицированного коэффициента Dice (modified Dice coefficient), вхождение правдоподобия (log likelihood ratio), оценку χ^2 (Chi-square test). Вместе с тем, без специальных модификаций никакой из этих методов не позволяет учитывать одновременно уровень позитивного и негативного отношения одного субъекта к ключевым вопросам, учитывать взаимную зависимость ключевых вопросов, вплоть до смысловой синонимии.

Применение модели определения репутации в социальных сетях на базе использования ГЧС может обеспечивать: возможность обучения системы, учет некоторой смысловой синонимии, омонимии в ключевых вопросах на уровне расширения таблиц умножения соответствующих ГЧС; возможность применения имеющихся наработок в области гиперкомплексных числовых систем, в том числе изоморфных ГЧС, но более пригодных для вычислений.

4.3. Рейтингование интернет-ресурсов

С проблемой управления репутацией в сети Интернет тесно связано понятие живучести информации. В свою очередь, для управления живучестью информационных объектов необходимо моделирование их жизненного цикла: формирования и развития, реакции на деструктивные воздействия, восстановления, разрушения.

Под живучестью информационной системы понимают способность ее (или ее фрагмента) адаптироваться к новым непредусмотренным условиям, противостояния нежелательным влияниям при одновременной реализации основной функции – целевого информирования. Кроме того, с живучестью информационных объектов сегодня связывают такая

социально важная проблема, как обеспечение информационной безопасности.

Существует несколько механизмов, обеспечивающих живучесть информационных объектов в Интернете.

Ниже рассматриваются некоторые наиболее распространенные механизмы обеспечения живучести, которые в реальности применяются не в чистом виде, а как правило, в комбинированном.

Для изучения проблем, связанных с живучестью необходимо четко определить как само это понятие, так и привести формальную модель, на основании которой можно рассчитывать уровень живучести для таких трудно формализуемых сущностей, как информационные объекты.

5.3.1. Механизмы обеспечения живучести информационных объектов

Понятие живучести информационной составляющей сети Интернет подразумевает способность информационных объектов (новостных сообщений, статей, документов, видеороликов и т.д.) своевременно выполнять свои функции (информирования) в условиях действия дестабилизирующих факторов. Такими факторами могут быть устранение отдельных объектов из информационного пространства, потеря ими свойств актуальности, доступности [Додонов, 2011], [Knight, 2003]. Рассмотрим некоторые из них.

1. Копирование данных при размещении их на целевой ресурс. То есть автор размещает информацию, которая копируется хостинг-провайдером на некоторое количество зеркальных серверов. Пример – скандально известная служба WikiLeaks (несколько сотен серверов, на которых хранятся фрагменты копий).

2. Перепечатка информации (републикации, «копипаст») на другие сайты с целью их информационного наполнения. В качестве примера приводится соотношение оригинальной информации и общего объема информации, сканируемой системой InfoStream [Григорьев, 2007] за первые четыре месяца 2012 г. по дням (рис. 87). При этом следует отметить, что наиболее важная и интересная информация перепечатывает-

ся сотни раз, в то время как неактуальная, неинтересная информация практически не дублируется.

3. Размещенная однажды информация навсегда попадает в архивные службы Интернета типа Архив Интернета (archive.org), который накапливает сетевую информацию. Библиотека Конгресса США (www.loc.gov) купила права на хранение всех публичных сообщений социальной сети Twitter с 2006 года и всех твиттов, которые будут опубликованы впредь. Библиотека Конгресса также реализует и национальный проект сохранения и распространения цифрового контента Digital Preservation (www.digitalpreservation.gov – 1400 коллекций данных).

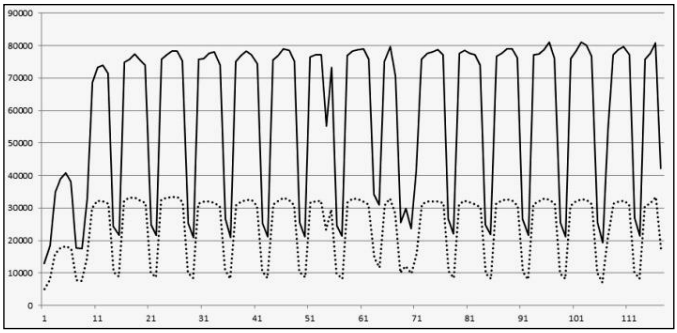


Рис. 87 – Соотношение оригинальной информации (пунктирная линия) и общего объема информации (сплошная линия), сканируемой системой InfoStream

4. Информация часто остается в кешах поисковых систем, даже если она удалена с веб-страницы или страницы социальной сети. Информация индексируется глобальными информационно-поисковыми системами и остается у них в кеш-памяти, откуда она доступна пользователям. Лишь относительно недавно у администраторов веб-ресурсов появилась возможность самостоятельного удаления своего контента из кешей Google и Яндекс. Часто многое, например, о человеке можно узнать в его блоге, онлайн-репутация – сегодня модный бренд. Что касается социальной сети Twitter, то twitFlink (www.twitflink.com), к примеру, который быстро соберет и выдаст твитты пациента. Сервис Google Replay позволяет

находить и просматривать тематические сообщения в микроблогах за указанный период времени.

5. И наконец, информация с веб-сайта может сохраняться на локальных компьютерах конечных пользователей, которые получили к ней доступ либо непосредственно, либо через интеграторов информации.

5.3.2. Формальные модели живучести информационных объектов

Известно, что живучесть информационного объекта можно оценивать как вероятность того, что объект будет неповрежденным в течение определенного периода времени t при определенных условиях [Li, 2012].

Если информационный объект сохраняется на n серверах (носителях информации), то вероятность разрушения этого объекта оценивается как:

$$F_{lost}(t) = \prod_{i=1}^{n} F_i(t).$$

В этом произведении $F_i(t)$ – вероятность потери информационного объекта на i-м сервере за время t.

Соответственно живучесть оценивается как:

$$S_n(t) = 1 - F_{lost}(t) = 1 - \prod_{i=1}^{n} F_i(t).$$

Допуская, что вероятность уничтожения информационных объектов пропорциональна времени их существования, и то, что время их разрушения имеет степенное распределение (в соответствии с законом Парето), можно считать целесообразным и обоснованным исследование модели со степенным распределением потерь информационных объектов, что принципиально отличается от подходов, в которых используется пуассоновский поток ошибок (теория систем массового обслуживания) и распределение ошибок по Вейбуллу. В этом случае, живучесть можно оценивать как:

$$S_n(t) = 1 - \prod_{i=1}^{n} F_i(t) = 1 - \prod_{i=1}^{n} Ct^{-\beta} = 1 - C^n t^{-n\beta},$$

где C, β – некоторые константы.

Закономерности статистического распределения времени жизни информационных объектов позволяют делать выводы, связанные с их живучестью, а именно учитывать явления самоподобия, нерегулярности потери информации, наличие «тяжелого хвоста» в распределении, которое характеризует чрезвычайно большое количество фактически устаревших информационных объектов и т.п.

При анализе жизненного цикла информационного объекта можно использовать еще два больших класса моделей: булевы и марковские.

В булевой модели можно предположить, что копии информационного объекта содержатся на n серверах, при этом i-му серверу соответствует булева переменная x_i, которая может принимать значения {0,1}, т.е. $x_i = 1$, если информационный объект на сервере i активен, и 0 – в противном случае. Состояние информационного объекта определяется структурной функцией его доступности (булевой функцией) $S(x_1, x_2, \ldots, x_n)$, которая принимает значения 1, если информационный объект доступен, и 0 в противном случае.

Если доступность информационного объекта рассматривать как функцию времени, то состояние информационного объекта на i-м сервере можно рассматривать как случайный процесс $x_i(t)$, принимающий в произвольные моменты времени $t \geq 0$ значения 0 и 1. Для системы определяется вероятность ее безотказной работы по приведенным выше формулам.

Среди недостатков булевых моделей можно назвать предположение только о двух состояниях информационного объекта – активности (доступности) и неактивности. Кроме того, в общем случае характер отказов отдельных копий информационного объекта зависит от состояния других копий.

Информационный объект можно описать также марковской моделью. Пусть система (множество копий информационного объекта) имеет m возможных состояний. Обозначим множество состояний через $M = \{z_1, z_2, \ldots, z_m\}$. Для любого фиксированного момента времени $t \geq 0$ состояние системы $z(t)$ интерпретируется как случайная величина. Заданы

множество всех состояний M, вектор распределения начальных вероятностей $p(0)$ и функция переходных вероятностей. Определяется вероятность активности, «жизни» системы в заданный момент времени t (готовность системы). Применимость марковских моделей также имеет свои границы. Интенсивности переходов между отдельными состояниями системы могут быть нестационарными, принимаемые при расчете допущения относительно распределения интенсивности отказов могут значительно снизить точность полученных результатов; число состояний системы может быть так велико, что расчет становится практически невозможным.

Оценка живучести информационных объектов может проводиться на всех этапах их жизненного цикла. Существует несколько подходов, к проведению оценки живучести, имеющих наиболее общий характер. Живучесть можно оценить относительно некоторого стандартного внешнего воздействия либо относительно множества внешних воздействий. В этом случае решается задача нахождения множества характеристических векторов состояний информационного объекта (в простейшем случае – распределение по серверам), в которых реализуется конфигурация, обеспечивающая выполнение цели функционирования. Мощность этого множества может служить мерой живучести всего информационного объекта.

При анализе живучести информационных объектов рассматривается проблема информирования по их различным аспектам, независимо от наличия или отсутствия неблагоприятных факторов. В связи с этим, в качестве количественного критерия оценки живучести целесообразно использовать отношение количества функций, выполняемых объектом при наличии определенных неблагоприятных воздействий либо множества таких воздействий, к общему количеству функций информационного объекта, с учетом критичности выполняемых и не выполняемых функций. Критичность каждой конкретной функции определяется индивидуально для каждого конкретного информационного объекта исходя из его специфики. Количественный показатель живучести конкретного информационного объекта в заданных условиях можно вычислять по формуле:

$$S = \sum_{i \in \Delta} \alpha_i \,/\, \sum_{j \in \Theta} \alpha_j \,,$$

где Θ – множество всех функций информирования, Δ – множество функций информационного объекта, выполняемых в заданных условиях $(\Delta \subseteq \Theta)$, α_n – критичность n-й функции. Таким образом, количественная оценка живучести информационного объекта **будет изменяется** в интервале [0, 1], живучесть тем выше, чем больше ее количественная оценка.

5.3.3. Цифровые следы и тени

Удаление информационного объекта с веб-ресурса не может гарантировать его исчезновения из Интернета. Остаются не только «цифровые следы» и «цифровые тени».

Выражение «цифровые следы» (Digital Footprint) относятся к той информации, которая оставляется самим пользователем при работе в Сети и по которой можно не только его идентифицировать, но и «привязать» к определенным действиям, событиям, восстановить какие-то фрагменты биографии.

Часто пользователи по доброй воле указывает свои Ф.И.О., «привязывая» дальнейшую информацию к собственной личности, дату рождения, семейное положение, образование, профессию, места предыдущей работы и много чего еще, включая и контактные телефоны, и адреса электронной почты. Кроме «цифровых следов», которые пользователи оставляют сами, информация о пользователях постоянно тиражируется и без всякого его участия.

Информация о пользователе, создаваемая без его участия, получила название «цифровой тени» (Digital Shadow), которые возникают и накапливаются всякий раз, когда кто-то ищет пользователя через поисковые системы, когда происходит электронная почтовая рассылка по спискам, в которых он фигурирует и во многих других случаях. Индексация роботами поисковых машин страниц с информацией пользователя и их последующее кеширование – это тоже создание «цифровой тени», доступной каждому. Кроме «цифровых теней открытого доступа», создаются и копятся «цифровые тени ограниченно-

го доступа» – записи камер наблюдения, банковские транзакции, биллинги интернет-магазинов, сервисов продажи билетов, телефонных звонков и др.

По оценке аналитической компании International Data Corporation (IDC), специализирующейся на исследованиях рынка информационных технологий, объем «цифровой тени», т.е. информации о пользователе Интернет, которая создается без его участия, уже в 2007 г. превысил объем информации, которую создает сам пользователь.

С проблемой репутации в Интернете ежедневно сталкивается все больше пользователей. Об этом свидетельствует и появление особых сайтов (например, www.suicidemachine.org), позволяющих одновременно удалить регистрацию и все сделанные записи на различных форумах и в социальных сетях. Такая операция называется «покончить с собой в Интернете». Однако эта система пока несовершенна. С недавнего времени эту заботу берут на себя специальные компании, так называемые «интернет-чистильщики», которые налаживают контакты с администрацией ведущих поисковых систем и социальных сетей, отдельных веб-сайтов, используют программные интерфейсы взаимодействия с кешами поисковых систем.

В качестве иллюстрации можно привести данные администрации социальной сети (сервиса микроблогов) Twitter о количестве запросов об удалении контента. По данным аналитиков, за первое полугодие 2013 года правительства разных стран направили в Twitter 1157 запросов о предоставлении информации. За аналогичный период 2012 г. эта цифра составляла 849. При этом в 10 раз выросло количество запросов об удалении контента. По числу запросов об удалении информации лидирует Россия. Кроме того, отмечается резкий рост правительственных запросов. 78 % всех запросов об информации (902) приходятся на долю США. На втором месте и третьем месте находится Япония (87) и Великобритания (26).

Понятие живучести информационного объекта подразумевает его способность своевременно выполнять свои функции (в данном случае – информирования) в условиях действия дестабилизирующих факторов. Такими факторами могут быть устранение отдельных информационных объектов из информационного пространства, потеря их актуальности, доступности. Необходимо отметить, что привлечение внима-

ния аудитории к другой теме, порождение другого информационного объекта также может снизить актуальность текущего информационного объекта.

При этом следует учитывать, что самая важная информация, попав в Интернет, остается там практически навсегда, и как показывает практика, рассчитывать на ее легкое удаление или изменение не приходится. Лучшим методом оказывается вытеснение нежелательной информации новыми сюжетами, проведение специальных мероприятий по содержательному исправлению ошибок [Додонов, 2010].

Учитывая эффект сверхживучести информации в сети Интернет, стоит учитывать несколько важных моментов, при борьбе с негативным контентом при управлении репутацией в сети:

- нельзя просто проигнорировать негативный контент; как известно, информационные сообщения, особенно негативной направленности, многократно дублируются в сети. Поэтому необходимы опровержения, позитивный контент;
- интернет-чистильщики – службы устранения негатива из сети Интернет могут «механически» лишь частично решить проблему. Негативная информация всеравно где-то останется и когда-то всплывет. Поэтому следует вытеснить негативный контент позитивным;
- позитивный контент должен быть правдивым, объективным. Интернет – отличный детектор лжи;
- необходимо размещать «выталкивающую негатив» позитивную информацию в сети на различных целевых ресурсах, заботясь о гиперссылках на эту информацию.

Живучесть информационных объектов и систем трудно заметить в нормальных условиях функционирования. Это свойство рельефно проявляется только в случаях потери информации, возникновения нарушений в структуре информационной системы, отказе ее составляющих, отдельных функций, целенаправленных деструктивных влияний. В зависимости от класса систем, их сложности, степени организованности, а также от выбранного уровня анализа свойство живуче-

сти может оцениваться как устойчивость, надежность, адаптивность, отказоустойчивость.

Наблюдаемый в настоящее время процесс в области интеллектуализации автоматизированных систем, перехода от простой обработки данных к процессам поддержки принятия решений требует новых подходов. Именно поэтому особое место занимают задачи, связанные с обеспечением живучести, как информационных систем, так и информационных объектов в сетевой среде.

5. Правовые вопросы конкурентной разведки

5.1. Конкурентная разведка в правовом поле

Безусловно, конкурентная разведка как сфера деятельности должна осуществляться в рамках правового поля государства. Основой для этого являются конституционные права на поиск, получение, передачу и использование информации во всех цивилизованных государствах. При этом следует отметить, что в ряде стран законодательство, ограничивающее деятельность по сбору и обработке информации, практически ставит конкурентную разведку под запрет.

Вместе с тем, в Украине «каждый имеет право свободно собирать, хранить, использовать и распространять информацию устно, письменно или любым другим способом – по своему выбору» (Конституция Украины, раздел 2, ст. 34).

Таким образом, в Украине правовое регулирование в информационной сфере, к которой, безусловно, относится и конкурентная разведка, основывается на следующих принципах:

1) свобода поиска, получения, передачи, производства и распространения информации любым законным способом;

2) установление ограничений к доступу информации только законами государства;

3) открытость информации о деятельности государственных органов и органов местного самоуправления и свободный доступ к такой информации, кроме случаев, установленных законами государства;

4) по категории доступа информация подразделяется на открытую (общедоступную) и с ограниченным доступом. В свою очередь, информация с ограниченным доступом по своей правовой природе также подразделяется на два вида: сведения, составляющие государственную тайну; конфиденциальная информация.

Несмотря на то, что конкурентная разведка сегодня является признанной сферой деятельности, узаконенного понятия «конкурентная разведка» в Украине сегодня не существует, хотя деятельность по сбору, хранению, обработке и распространению информации регулируется целым рядом законодательных и нормативных актов:

Закон України «Про інформацію» від 02.10.1992 р. № 2657-XII (зі змінами від 13.01.2011 р.), ст. 5-7 .

Закон України «Про друковані засоби масової інформації (пресу) в Україні» від 16.11.1992 р № 2782-XII, ст. 6, 25.

Закон України «Про охоронну діяльність» № 4616-VI від 22.03.2012 р. ст. 9, 13, 19.

Закон України «Про захист персональних даних» № 2297-VI від 01.06.2010 р.

Цивільний кодекс України (ст. 505), Кримінальний кодекс України (ст. 231, 232), Кодекс України про адміністративні правопорушення (ст.163, ст.163);

Указ Президента України «Питання європейської та євроатлантичної інтеграції» від 20.04.2019 р. № 155/2019.

Указ Президента України «Про Національний Координаційній центр кібербезпеки» від 07.06.2016 р. № 242/2016.

Нельзя забывать, что осуществление мероприятий по обеспечению безопасности бизнеса даже в рамках конкурентной разведки иногда может быть воспринято как проведение оперативно-розыскной деятельности, проводить которую, согласно Закону Украины “Об оперативно-розыскной деятельности” от 18.02.1992 г. № 2135-XII могут лишь субъекты, указанные в отдельных статьях данных Законов. При этом перечень субъектов является исчерпывающим, а проводить оперативно-розыскную деятельность другими юридическими и физическими лицами запрещается.

В утвержденной Указом Президента Украины №96/2016 от 27 января 2016 года Стратегии кибербезопасности Украины декларируются основные задачи силовым органам, среди которых: «на разведывательные органы Украины – реализация разведывательной деятельности по выявлению угроз национальной безопасности Украины в киберпространстве, других событий и обстоятельств, касающихся сферы кибербезопасности», а также предусматривается «создание системы своевременного выявления, противодействия и нейтрализации киберугроз, в том числе с привлечением волонтерских организаций», всё это, безусловно, относится к применению средств OSINT (или конкурентной разведки) в этой обрасти.

В то же время действующими Уголовным кодексом Украины предусмотрено уголовная ответственность за незаконный

сбор с целью использования или использование сведений, составляющих коммерческую тайну, а также за разглашение коммерческой тайны. Очевидно, такие сведения выходят за рамки конкурентной разведки.

При достаточно широком толковании норм законодательства любые процедуры сбора, обработки и хранения информации о конкурентах становятся, с одной стороны, легитимными, практически безнаказанными, а, с другой стороны, затруднительными. В Украине фактически закрыт доступ к большому пласту свободно доступной в большинстве стран бизнес-информации, например, о недвижимости (имеющейся и заложенной), земельных участках, наличии банковских счетов и т.п. В этих странах большую часть сведений можно получить только путем консультаций с соответствующими экспертами.

Сегодня как никогда остро стоит проблема криминализации отдельных служб конкурентной разведки. Многие службы безопасности сегодня пользуются базами данных с информацией о персонах. Такие базы используются с вполне благими целями, например, для проверки данных о сотртниках, партнерах и конкурентах. Очевидно, такими базами данных они будут пользоваться и в дальнейшем, однако будут вынуждены нарушать закон и «уходить в подполье». Технически возможности использования и ведения подобных баз данных предоставляют многочисленные системы типа Cronos (оболочки, распространяемые вполне легально). С помощью подобных инструментальных средств любому заинтересованному пользователю сети Интернет становятся доступны многочисленные базы данных, работающих с этими оболочками.

В результате к деятельности компаний, занимающихся конкурентной разведкой, наблюдается повышенное внимание со стороны государственных контролирующих органов.

Это связано с несколькими группами правовых проблем, которые можно сгруппировать, выделив проблемы, связанные с:

1) защитой коммерческой тайны;

2) защитой персональных данных;

3) соблюдением авторских прав;

4) возможностью конкуренции на рынке самой конкурентной разведки.

Также можно выделить три класса основных проблем авторского права, имеющих отношение к конкурентной разведке, это проблемы, связанные с такими аспектами:

- правомерностью использования входной информации (источников информации), на основании которой формируются отчеты – результаты конкурентной разведки; проблемы,
- авторскими правами на результаты конкурентной разведки;
- правами на применение (использование) специализированного программного обеспечения, необходимого для проведения конкурентной разведки.

Кроме того, одна из проблем, стоящая перед службами конкурентной разведки в Украине – практически полное отсутствие антидемпингового законодательства. Несмотря на то, что приход на этот рынок крупных международных игроков затруднен ввиду отсутствия необходимых связей, баз данных, архивов и даже лингвистической и правовой подготовки, с их стороны возможно проявление демпинга на услуги конкурентной разведки.

Ситуация может измениться, если будет создана четкая правовая база для деятельности, связанной со сбором и аналитической обработкой информации и, в частности, для конкурентной разведки.

5.2. Конкурентная разведка и защита коммерческой тайны

Важное значение для становления конкурентной разведки имел ряд статей Закона Украины «О защите от недобросовестной конкуренции» № 236/96-ВР от 07.06.1996, где (ст. 15-1), запрещается «Неправомерный сбор коммерческой информации», «Разглашение коммерческой информации» «Неправомерное использование коммерческой информации» (гл 4, ст. 16, 17, 19, соответственно).

В постановлении Кабинета Министров Украины от 9 августа 1993 года № 611 «О перечне сведений, которые не составляют коммерческой тайны» определен целый класс документов, касающихся деятельности бизнес-структур, которые

являются фактически открытыми, в частности, учредительные документы, формы отчетности, информация об участии учредителей и должностных особ в других компаниях и т.п.

Зачастую усилия конкурентной разведки направлены на получение коммерческой тайны конкурентов. И хотя в различных законодательных актах даются различные формулировки, можно согласиться с тем [Іващенко, 2006], что коммерческая тайна характеризуется такой совокупностью признаков: информация является секретной, является неизвестной и не является легкодоступной для лиц, которые обычно имеют дело с видом информации, к которой она относится; в связи с тем, что является секретной, она имеет коммерческую ценность. Таким образом, коммерческая тайна – это информация, которая является полезной и не является общеизвестной обществу. Она имеет действительную или коммерческую ценность, с которой можно иметь прибыль и для защиты которой владелец принимает меры во всех сферах жизни и деятельности». Таким образом, можно сказать, что деятельность конкурентной разведки иногда направлена на добычу информации, которая не является общедоступной и охраняется законом. Эти деяния нарушают огромное количество статей Уголовного Кодекса Украины, в частности, статью 231 «Незаконный сбор с целью использования или использование сведений, составляющих коммерческую или банковскую тайну».

Таким образом, коммерческая разведка может легитимно использовать лишь те методы и способы сбора и обработки информации, которые не противоречат законодательству, т.е. основные функции конкурентной разведки — качественный сбор, систематизация и, главное, анализ информации, а не слежка, подкупы и незаконные хакерские взломы.

Впервые право на сохранение коммерческой тайны было провозглашено Законом СССР от 4 июня 1990 г. «О предприятиях в СССР». В ст. 33 указанного Закона раскрывалось понятие коммерческой тайны как не являющихся государственными секретами сведений, связанных с производством, технологической информацией, управлением, финансами и другой деятельностью предприятий, разглашение (передача, утечка) которых может нанести ущерб их интересам.

В настоящее время украинское законодательство об охране служебной и коммерческой тайны представляет собой совокупность статей, которые содержатся в различных правовых актах, посвященных в целом регулированию иных общественных отношений.

Гражданский кодекс Украины, в свою очередь, определяет коммерческую тайну (с. 505 п. 1) как информацию, «которая является секретной в том смысле, что она в общем или в определенной форме и совокупности является неизвестной и не легкодоступной для лиц, которые обычно имеют дело с видом информации, к которому она относится, в связи с этим имеет коммерческую ценность и была предметом адекватных существующим обстоятельствам мер, относящихся к сохранению ее секретности, предпринятых особой, которая законно контролирует эту информацию».

В соответствии с этими определениями, как только информация в результате каких-либо действий попадает, например, на страницы любого веб-сайта, она перестает считаться коммерческой тайной, так как становится легкодоступной.

Хотя во многих статьях Уголовного кодекса Украины (ст. 231, 232, 232-1, 361, 363) установлена уголовная ответственность как за разглашение коммерческой тайны, так и за незаконный сбор и использование сведений, к ней относящихся, однако, существующая нормативно-правовая база четко не регламентирует, какие именно сведения о финансово-хозяйственной деятельности предприятия являются коммерческой тайной (за исключением разве что банковской тайны, определение которой дано в ст. 60 Закона Украины «О банках и банковской деятельности»).

5.3. Конкурентная разведка и защита персональных данных

Государственные учреждения, банки, крупные корпорации не всегда могут обеспечить защиту хранящихся у них баз персональных данных, в результате чего, огромный поток конфиденциальной информации поступает на рынок. Обеспечение безопасности персональных данных – объективная потребность. Сегодня персональные данные, информация о людях превращается в самый дорогой товар. Такая инфор-

мация в руках злоумышленника – мощное оружие. То есть персональные данные необходимо защищать.

Персональные данные – важная составляющая более широкого понятия – приватность. Поэтому защита персональных данных, это составляющая часть обеспечения приватности. Приватность, наряду со свободой слова и другими правами, является одной из основных ценностей человечества.

На сегодня, основными европейскими документами в области защиты персональных данных являются Конвенция Совета Европы «О защите личности в связи с автоматической обработкой персональных данных» и Директива Европарламента «О защите физических лиц при автоматизированной обработке персональных данных», ETS № 108, 1981 г., которая является обязательной для всех государств-членов Европейского союза и которая является предметом для подражания в области законодательства, в том числе, и нашей страной. Страны Евросоюза последовательно приводят свое законодательство в соответствие с Директивой. В Великобритании еще в 1998 году был принят «Закон о защите персональных данных» – «Data Protection Act 1998». Его техническая реализация – проект стандарта «Specification for the management of personal information in compliance with the Data Protection Act 1998» (BS 10012, 2009). Параллельно с англичанами свою версию стандарта по безопасности персональных данных выпустили в США. Проект документа по защите персональных данных для американских государственных структур – «Guide to Protecting the Confidentiality of Personally Identifiable Information (PII)» (SP 800122) регламентирует выполнение Законов «The Privacy Act of 1974» и «Privacy Protection Act of 1980». Канада выпустила «Privacy Code» – набор документов для реализации законодательства по защите сведений о частных лицах (The Privacy Act и PIPEDA).

В государствах-членах Евросоюза определения персональных данных, как правило, максимально широкие, в результате чего гражданами на практике зачастую не выполняется соответствующее законодательство из-за излишней «нагрузки». Соответствующие органы государственной власти, как правило, не предпринимают никаких действий, кроме особых случаев. Важными остаются вопросы возникновения коллизий между требованиями приватности и интересами свободы

слова. Современными европейскими законами, как правило, запрещается сбор, хранение, использование и распространение без согласия субъекта данных именно критичных персональных данных.

Право на приватность гарантируется Конституцией Украины. Статья 32 Конституции Украины гласит: «Никто не может подвергаться вмешательству в его личную и семейную жизнь, кроме случаев, предусмотренных Конституцией Украины». Кроме того в Конституции Украины предусмотрена защита еще некоторых аспектов приватности. Так, статья 30 защищает неприкосновенность жилища (территориальная приватность), статья 31 – тайну переписки, телефонных разговоров, телеграфной и другой корреспонденции (коммуникационная приватность), статья 32 предусматривает запрет сбора, хранения, использования и распространения конфиденциальной информации о лице без его согласия (информационная приватность), а статья 28 предусматривает запрет подвергать лицо без его свободного согласия медицинским, научным или другим исследованиям (защищая некоторые элементы физической приватности).

Конвенция о защите лиц в связи с автоматизированной обработкой персональных данных Страсбург, 28 января 1981 (ратификация от 06.07.2010) определяет положение относительно передачи через национальные границы с помощью любых средств персональных данных, подвергающихся автоматизированной обработке или собранных с целью их автоматизированной обработки.

Следующие данные часто используются для выделения конкретного лица, указанные как личные Управлением США по менеджменту и бюджету:

- полное имя (имеется в виду имя вместе с фамилией)
- национальный идентификационный номер;
- IP-адрес (в некоторых случаях);
- номерной знак транспортного средства;
- номер водительских прав;
- лицо, отпечатки пальцев, или почерк;
- номера кредитных карт;
- цифровая идентичность (цифровая подпись);
- дата рождения;
- место рождения;

- генетическая информация.

Согласно законодательству большинства европейских государств персональные данные разделяются по критерию «чувствительности» на данные общего и «чувствительные» (уязвимые) личные данные.

Общие личные данные:

- идентификационные данные (фамилия, имя, отчество, адрес, телефон и т.п.);
- паспортные данные;
- личные сведения (возраст, пол, семейное положение и т.д.);
- состав семьи;
- образование;
- профессия;
- жилищные условия;
- образ жизни;
- жизненные интересы и увлечения;
- потребительские привычки;
- финансовая информация.

«Чувствительные» личные данные:

- информация о расовом, этническом происхождении и национальности;
- сведения, касающиеся политических, мировоззренческих и религиозных убеждений;
- сведения о членстве в политических партиях, профсоюзах, религиозных или общественных организациях;
- сведения о состоянии здоровья и половую жизнь;
- генетические и биометрические данные;
- место нахождения и пути передвижения лица;
- информация о применении к лицу мер в рамках трудового следствия;
- информация о совершении в отношении лица различных видов насилия.

Конституционные нормы определяют исчерпывающий перечень оснований для вмешательства в приватность и условий для такого вмешательства. Однако в постсоветских государствах существует много отраслевых норм права, противоречащих требованиям их Конституций. Именно такие нормы

не соответствуют международным стандартам, практике Европейского законодательства.

В соответствии с украинским законодательством персональными данными в Украине является Ф.И.О. в сопровождении любых других идентификационных данных, например, адреса, телефона или образовательного статуса.

Для выяснения, какое же отношение имеет физическое лицо или компания к защите персональных данных, большое значение имеет определение субъектов отношений, связанных с персональными данными (статья 4 Закона Украины № 2297-VI): «Субъектами отношений, связанными с персональными данными, являются:

- субъект персональных данных;
- владелец базы персональных данных;
- распорядитель базы персональных данных;
- третье лицо;
- уполномоченный государственный орган по вопросам защиты персональных данных;
- другие органы государственной власти и органы местного самоуправления, к полномочиям которых относится осуществление защиты персональных данных.

В украинском законодательстве предусмотрен уведомительный характер обработки персональных данных. Владелец или распорядитель (оператор) до начала обработки персональных данных обязан уведомить уполномоченный орган по защите прав субъектов персональных данных о своем намерении осуществлять обработку персональных данных. Затем данные о владельцах или распорядителях (операторах) вносятся в специальный реестр операторов. Информация, содержащаяся в реестре операторов, становится общедоступной.

Законы о персональных данных касаются большинства населения как участников процесса «обработки» данных. А так как субъектом персональных данных является каждый человек, то Закон носит всеобщий характер и касается каждого.

Этот законодательный акт имеет прямое отношение к сфере информационных технологий и телекоммуникаций, оба содержат спорные, противоречащие сложившейся практике, казалось бы, неосуществимые положения. Требования закона распространяются на все юридические и физические лица, и интернет-сфера не является исключением. Закон о защите персональных данных может изменить принципы работы украинских интернет-ресурсов: сервисов электронной почты, знакомств, онлайн-магазинов и социальных сетей, хотя сами участники рынка надеются, что сайты не подпадут под действие закона. Владельцам интернет-ресурсов для соблюдения всех положений закона о персональных данных необходимо тщательным образом продумывать организацию своей деятельности. В настоящее время существует немало веб-служб, в рамках, которых происходит сбор, хранение, использование персональных данных. Соблюдение требований закона является непростой задачей для владельцев этих ресурсов, в частности, чиновники имеют возможность обязать интернет-компании брать письменное согласие на использование анкетных данных у каждого пользователя. Не секрет, что на многих сайтах размещается информация, содержащая персональные данные людей (например, ищущих работу, знакомства), в том числе и относящиеся к специальным категориям, например, национальность или вероисповедание. Задача тех, кто обеспечивает подобные сервисы, легитимно обрабатывать подобную информацию и одновременно защищать ее согласно требованиям законодательства.

В частности, персональные данные широко используются в социальных сетях и сервисах электронной почты. Например, владельцам веб-ресурсов весьма сложно соблюсти требование закона о получении согласия каждого пользователя на обработку его персональных данных. При этом закон возлагает именно на оператора обязанность доказывания факта получения им такого согласия.

Современная интернет-компания собирает и обрабатывает разные категории персональных данных – своих сотрудников, своих контрагентов по договорам и некоторые данные пользователей своих сервисов. Люди, размещающие информацию о себе в социальных сетях или службах знакомств, сознатель-

но делают ее открытой для всех пользователей ресурса, и по закону ее можно трактовать как «общедоступную», а значит, соблюдения особого режима конфиденциальности в отношении ее не требуется, но в социальных сетях есть и информация, которую пользователь скрывает, делая ее доступной только для отдельной группы пользователей («друзей»). В этом случае интернет-ресурс должен предусматривать для нее специальные средства защиты.

В практике конкурентной разведки приходится сталкиваться с многочисленными противоречиями и казусами в существующем законодательстве, например, в украинском Законе «О защите персональных данных» (часть 9 ст. 6) говорится: «использование персональных данных в исторических, статистических или научных целях может осуществляться только в обезличенном виде». То есть записи в отчетах конкурентной разведки должны выглядеть примерно так: «Персона А провела переговоры с персоной Б». В научных отчетах нельзя делать ссылок на других коллег, даже при наличии их письменного согласия. Вызывает определенные сложности и необходимость оповещать орган власти «о каждом изменении сведений, необходимых для регистрации соответствующей базы», которая среди прочего включает информацию обо всех распорядителях (пользователях) такой базы данных.

Кроме того, многие службы конкурентной разведки, совершенно на законных условиях создающие базу данных персональных данных для решения обозначенной ими задачи, обязаны уничтожить плоды своей работы, достигнув цели. А ведь, если основная цель, например, при оказании услуг клиентам – это выполнение этих самих заявок, но сопутствующая цель любой уважающей себя организации – это и наработка базы клиентов. И эта база часто имеет собственное коммерческое значение. Известны многочисленные случаи легальной перепродажи баз данных клиентов, например, при прекращении деятельности фирмы-владельца. В украинском законодательстве строгой статьи нет, однако предусмотрены условия уничтожения персональных данных, среди которых (ст. 15), «прекращение правоотношений между субъектом персональных данных и владельцем или распорядителем базы…». А это означает, например, что оператор – исполнитель

услуги должен уничтожить всю наработанную за время выполнения услуги базу данных.

Поэтому владельцы и распорядители подобных баз данных переформулируют свои цели специальным образом, например, как «оказание услуги с возможностью хранения персональных данных в течение гарантийного срока…». Таким образом, соблюдаются нормы законодательства и обеспечиваются интересы исполнителя – владельца или распорядителя (оператора) базы персональных данных.

Подразделения конкурентной разведки занимаются обработкой персональных данных, которые находятся в открытых источниках в сети Интернет, т.е. являются общедоступными. Для их обработки согласия субъекта персональных данных не требуется. Однако при этом обязанность доказательства, что обрабатываемые персональные данные являются общедоступными, возлагается на владельца или распорядителя. А это значит, что необходимо либо накапливать доказательства того, что данные взяты из общедоступных источников, либо получать согласие от субъекта персональных данных и затем хранить этот документ. Кроме того, нужно иметь документ, подтверждающий общедоступность источника персональных данных. При этом остается без ответа вопрос доказательства того, что владелец информационного ресурса (веб-сайта) обладает письменным согласием на обработку.

Как никогда острой стала проблема криминализации отдельных служб конкурентной разведки. Многие службы безопасности сегодня пользуются базами данных с информацией о персонах. Такие базы используются с вполне благими целями, например, для проверки данных о сотрудниках, партнерах и конкурентах. Очевидно, такими базами данных они будут пользоваться и в дальнейшем, однако будут вынуждены нарушать закон и «уходить в подполье». Технически возможности использования и ведения подобных баз данных предоставляют многочисленные системы типа Cronos (оболочки, распространяемые вполне легально). С помощью подобных инструментальных средств любому заинтересованному пользователю Интернет становятся доступны многочисленные базы данных, которые работают под этими оболочками.

На государственном уровне в США основным правовым механизмом ведения разведки в открытых источниках мини-

стерства обороны является Совет по защите открытых источников (DOSC). Он служит форумом для координации и содействия мероприятиям и программам ведения разведки в открытых источниках для всех служб и боевых команд. Данный совет консультирует и докладывает заместителю министра обороны по разведке о вопросах ведения разведки в открытых источниках, о новых инициативах по улучшению эффективности работы подразделения OSINT и деятельности министерства обороны в целом. В обязанности Совета входят:

- координирует деятельность подразделения OSINT и утверждает его план ведения разведки в открытых источниках;
- определяет последовательность требований к процессу ведения разведки в открытых источниках.

Армейский стандарт США «ATP 2-22.9» устанавливает общие понятия, основные концепции и методы сбора разведывательных данных из открытых источников для Армии США. В этом документе подчеркивается характеристикуа OSINT как разведывательной дисциплины, его связи с другими разведывательными дисциплинами, и возможности его применения в ходе объединенных операций.

Использование общедоступной информации является важным аспектом технической разведки (TECHINT). Несмотря на то, что намерения, возможности и факторы уязвимости противников и потенциальных угроз подлежат засекречиванию, результаты OSINT (в частности, открытого сервиса «Google Earth») способствуют получению информации о самых скрытных государствах и организациях. Такие примеры свидетельствуют об ответственности деятельности в этой области.

Авторское право является одной из форм защиты, опубликованных и неопубликованных работ, предусмотренных главой 17 Кодекса США, что определяет авторов «оригинальных работ авторов», в том числе литературных, драматических, музыкальных и художественных произведений.

Национальные законы об авторских правах являются ограничениями конкурентной разведки. Нарушение прав, в частности, предусмотренных главой 17 Кодекса США, законами об авторских правах, все же оставляют возможность

правомерного использования конкурентной разведки, что определяется четырьмя факторами:

- целью и характером использования;
- свойствами, используемых авторских работ;
- количеством и частями авторской работы, которые используются;
- воздействием использования авторских работ на потенциальный рынок или ценность этих работ.

5.4. Конкурентная разведка и защита авторского права

Можно выделить три класса основных проблем авторского права, имеющие отношение к конкурентной разведке, это проблемы, связанные с такими аспектами:

- правомерностью использования входной информации (источников информации), на основании которой формируются отчеты - результаты конкурентной разведки;
- проблемы с авторскими правами на результаты конкурентной разведки;
- права на применение (использование) специализированного программного обеспечения, необходимого для проведения конкурентной разведки.

Кроме того, одна из проблем, стоящих перед службами конкурентнои разведки в Украине – практически полное отсутствие антидемпингового законодательства:

- ситуация может измениться, если будет создана четкая правовая база для конкурентной разведки;
- авторское право является одной из форм защиты, опубликованных и неопубликованных работ, предусмотренных главой 17 Кодекса США, что определяет авторов «оригинальных работ авторов», в том числе литературных, драматических, музыкальных и художественных произведений. Национальные законы об авторских правах являются ограничениями конкурентной разведки. Несмотря на это, все же остаются возможности правомерного использова-

ния конкурентной разведки, определяется четырьмя факторами:

- целью и характером использования;
- свойствами авторских работ;
- количеством и частями авторской работы;
- влиянием использования авторских работ на потенциальный рынок или ценность этих работ.

6. Противодействие информационным операциям

В последние годы благодаря многочисленным документам и публикациям Министерства обороны США стал популярен термин «информационные операции», прежде всего потому, что информационные технологии играют постоянно увеличивающуюся роль в военных операциях. При этом информационные операции определяются как «акции, направленные на воздействие на информацию и информационные системы противника, и защиту собственной информации и информационных систем» [DoD, 2003]. Информационные операции рассматриваются как объединение основных возможностей радиоэлектронной войны, компьютерных сетевых операций, психологических операций, военных действий и операций по обеспечению безопасности с целью воздействовать, разрушать, искажать информацию, необходимую для принятия противником решений, а также защищать собственную информацию.

Информационные операции охватывают целый комплекс процессов, проводимых в самых разных областях. При этом необходимо отметить, что информационные операции – существенная и традиционная составляющая боевых операций. Несмотря на то, что формальное определение в документах Департамента обороны США ориентировано на военные аспекты информационных операций, оно вполне применимо практически для любой области жизни.

Ниже будут рассматриваться такие информационные операции, которые реализуются с помощью информационных систем (ИС). Живучесть этих ИС во многом определяет живучесть информационных операций, которые реализуются в виде информационных воздействий на сознание людей.

Информация является отражением вложенного в нее смысла, поэтому сегодня информация превратилась из абстрактного термина в объект, цель и средство информационных операций, стала критическим понятием в проблематике безопасности. Бывший министр обороны США Уильям Коэн 18 марта 1999 г. заявил, что «способность армии использовать информацию, чтобы доминировать в будущих сражениях,

даст США новый ключ к победам в течение многих лет, если не в течение нескольких поколений» [Hill, 2000].

При моделировании и проведении информационных операций необходимо учитывать значение ценности информации для лиц, принимающих решение. Ценность информации включает ее своевременность, точность и «аналитичность». С практической точки зрения ценность информации также может быть определена как ее значимость или применимость, пригодность к использованию. Под применимостью информации понимается обеспечение доступа ЛПР к готовой к использованию информации. Стандарт ISO 9241 (ISO – Международная Организация по Стандартизации) определяет применимость в терминах эффективности и удовлетворения потребностей указанного набора пользователей для решения указанного набора задач в специфическом окружении. На практике большая часть полезной информации поступает к ЛПР от информационно-аналитических систем, обеспечивающих ориентацию в ситуации и поддержку при принятии решений. Согласно полевому уставу военного ведомства США «Информационные операции» (FM 100-6), «ориентация в ситуации означает комбинацию ясного представления о диспозиции своих и вражеских сил с оценкой ситуации и намерений со стороны командования».

Информационные операции осуществляются в некоторой социальной среде, соответственно, для успешного их проведения необходимо адаптироваться к этой среде, преодолеть определенный барьер не очень сильного внимания к информационному воздействию. Этот барьер возникает благодаря так называемой иммунной системе среды, которая может не пропустить информационные воздействия, если она достаточно мощная и/или уже научилась защищаться от подобных воздействий. К подготовительным действиям для проведения информационных операций может относиться создание «иммунодефицита» социальной среды путем воздействия через информационное пространство, например, с помощью материалов в СМИ. Очень часто информационные воздействия используют механизмы «вирусного маркетинга», например, в виде слухов, когда сенсационно поданная дезинформация распространяется с огромной скоростью. Именно иммунная система оказывает противодействие подобным информаци-

онным операциям. Очень часто с иммунной системой общества отождествляют государство, призванное обеспечивать безопасность этого общества, т.е. при наличии сильного государственного аппарата вероятность успеха антиобщественных информационных операций существенно снижается. Читатель прекрасно знает, как происходило противодействие подобным информационным процессам в тоталитарных государствах. В демократическом обществе, естественно, тоталитарные методы не применимы. В этом случае иммунитет достигается за счет «обучения», т.е. демократическое общество должно пройти через многие информационные атаки, воздействия, влияния стереотипов, чтобы выработать необходимый иммунитет.

Уровень готовности к проведению информационных операций сегодня считается ключевым фактором успеха проведения любой социальной процедуры, кампании.

Особой целью при проведении информационных операций являются информационно-аналитические системы субъекта воздействия. Оказывая влияния на такие системы, можно добиться того, что принимающие решение лица из лагеря противника примут неадекватные выводы, и требуемый социальный процесс изменит траекторию в необходимом оказывающей влияние стороне направлении [Горбулін, 2009] (рис. 88).

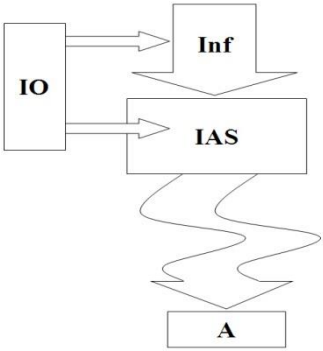


Рис. 88 – Воздействие на информационно-аналитическую систему противника: Inf – информационное пространство; IAS – информационно-аналитическая система; А – абонент системы – ЛПР; IO –ин информационные воздействия

В данном случае к непосредственным информационным воздействиям может быть отнесено размещение в информационном пространстве документов, компрометирующих противоположную сторону, реклама (в том числе скрытая) своих преимуществ, искаженные данные о внешней среде, искаженная информация о намерениях и т.д.

Социальные процедуры и процессы, как правило, сложно оценивать и моделировать, так как их результаты относятся к психологическим и социологическим, а не физическим. Именно этот факт также определяет проблематичность прогнозирования результатов моделирования информационных операций. Кроме того, экспериментирование с информационными воздействиями в рамках информационных операций более сложны и опасны, чем при моделировании физических процессов.

Действия для достижения эффективности влияния на процессы принятия решения противником иногда необходимо предпринимать в течение длительного времени, прежде чем они вступят в силу.

Одна из основных компонент информационных операций – социальное влияние, охватывающее все многообразие процессов влияния. Существенные изменения в убеждениях или отношении людей к некоторой проблеме или явлению, как ожидается, будут вести к изменению в поведении, связанном с этой проблемой.

В 1948 году Харольд Д. Лассвел [Lasswell, 1948] разработал модель трансмиссии коммуникаций, состоящую из пяти компонент:

- источник – персона, которая влияет или убеждает другие персоны;
- сообщение – с помощью чего источник пробует убедить цель;
- цель – человек, на которого источник пробует влиять;
- канал – метод доставки сообщений;
- воздействие — реакция цели на сообщение.

Хотя Лассвел прежде всего интересовался массовой коммуникацией, его модель передачи информации может применяться в межличностной коммуникации типа циркулярных моделей Шеннона–Вивера (Shannon–Weaver) и Осгуда–

Шрамма (Osgood–Schramm), которые включают петли обратной связи в процессе коммуникаций, утверждая, что коммуникация является циркулярным, а не линейным процессом [Schramm, 1974], [Osgood, 1954].

Моделирование объективных факторов социального влияния требует междисциплинарных подходов, имеющих отношение к информатике, маркетингу, политологии, социальной психологии. Самые известные модели формирования общественного мнения и социального влияния базируются на теории Латэйна динамического социального воздействия [Latane, 1981], [Latane, 1997], развитой многими другими авторами, прежде всего, в работах [Nowak, 1990], [Lewenstein, 1993], [Kacperski, 2000], [Sobkowicz,2003].

Пытаясь обосновать механизм социального влияния сообщений Латэйн [Latane, 1981] подчеркнул важность трех признаков отношений источника и цели:

- сила – социальная сила, вероятность или уровень влияния на индивидуумов;
- непосредственность – физическое или психологическое расстояние между индивидуумами;
- число источников – количество источников, стремящихся к цели.

Современное состояние моделирования информационных операций характеризуется рядом открытых проблем, основные из которых относятся к пониманию понятий информационного влияния и воздействия.

6.1. Информационное влияние, атаки и операции

Универсальными характеристиками объектов являются его состояние и возможность воздействия на другие объекты. Реализация возможности воздействия требует определенных условий, которые принято называть его влиянием. При этом объект, который может осуществлять свою волю, называют субъектом, а управлением принято называть воздействие по отношению к объекту воздействия, применяемое с определенной целью.

Когда индивидуум является целью влияния одного или более источников, динамическая социальная теория воздействия утверждает, что уровень социального влияния на индивидуума может быть представлен уравнением, являющимся

основой так называемой индивидуум-ориентированной модели:

$$I_i = -S_i\beta - \sum_{j=1, j\neq i}^{N} \frac{S_j O_j O_i}{d_{i,j}^{\alpha}},$$

где I_i — величина (количество) социального давления, оказываемого на индивидуума i, $(-\infty < I_i < \infty)$; O_i и O_j представляет мнение индивидуума (i и j, соответственно) по актуальному вопросу – +1 или –1 – поддержку или возражение относительно данного вопроса, соответственно. S_i (S_j) представляет силу индивида i (j) или влияние ($S_i > 0,\ S_j > 0$); β – сопротивление индивидуума к изменениям ($\beta > 0$); $d_{i,j}^{\alpha}$ – расстояние между индивидуумами i и j ($d_{i,j}^{\alpha} \geq 1$); α – показатель сокращения расстояния ($\alpha \geq 2$); N – общее количество агентов (индивидуумов, составляющих сообщество). Значение β, тенденция сохранять собственное мнение или сопротивляться изменению определяет то, что индивидуумы в рамках модели могут требовать больших или меньших объемов социального давления для изменения их мнения. Большие уровни значения α соответствуют эффекту возрастания расстояния между источником и целью, что влияет на объем социального давления на цель.

На основе введенных терминов формулируется понятие «информационного поля объекта» [Кононов, 2003], описываются его характеристики. Это дает возможность определить информационное воздействие как воздействие на информационное поле объекта. Исследуя информационные поля объектов и субъектов социальных систем, можно определить информационные влияния и управления. При этом информация может рассматриваться и как объект, и как средство воздействия. Использование информации как средства воздействия требует в процессе управления осуществить подготовку данных, производство соответствующей информации, а лишь затем реализовывать созданную информацию в виде воздействия (влияния).

Одним из основных методов ведения информационных операций является информационное влияние, оказываемое с целью информационного управления. Под информационным управлением в данном случае понимается механизм управления, когда управляющее воздействие носит неявный, косвенный информационный характер и объекту управления дается определенная информационная картина, под влиянием которой он формирует линию своего поведения. Таким образом, информационное управление — это способ воздействия, побуждающий людей к упорядоченному поведению, выполнению требуемых действий.

В соответствии с [Кононов, 2003], [Кульба, 2004] процесс информационного влияния одного объекта на другие целесообразно декомпозировать на следующие этапы:

- генерация источником влияния данных, информационных элементов и информационных совокупностей;
- передача информации источником влияния;
- прием информации реципиентом;
- генерация совокупности данных, информационных элементов и новых совокупностей объекта влияния;
- соответствующие активные действия объекта влияния.

Информационные воздействия на элементы систем можно классифицировать по таким признакам, как источники возникновения, длительность воздействия, природа возникновения и т.п.

Для выбора конкретных способов реализации информационного управления необходимо конкретизировать задачи, решаемые с помощью информационного воздействия, провести анализ процесса формирования информационных операций и выработать критерии их оценки. Информационное управление рассматривают как процесс, охватывающий такие три взаимосвязанных направления:

- управление обменом данными между реальным миром и виртуальным миром субъекта влияния;
- управление виртуальным миром субъектов влияния, механизмами принятия решений;
- управление процессом преобразования решений в действия субъектом влияния в реальном мире.

Информационное воздействие может быть двух основных видов:

1) изменение в требуемую сторону данных, которые использует информационно-аналитическая система объекта воздействия при принятии решений;

2) непосредственное влияние на процесс принятия решения объекта воздействия, например, на процедуры принятия решения или отдельные лица, принимающие решения.

Важнейшее значение для проведения информационных операций имеет окружающая среда, состояние объектов информационного воздействия, их взаимное влияние. В частности, если в качестве объектов информационных операций выбирается некоторое электоральное поле, то важно учитывать все электоральные популяции, входящие в это поле, которые представляют сторонников (или противников) тех или иных политических сил. Несмотря на то, что в дальнейшем будут рассматриваться и некоторые модели, в которых в явном виде постулируется однородность среды, в общем случае по отношению к информационным операциям окружающая среда может состоять из областей:

- доминирующего восприятия;
- повышенной чувствительности;
- индифферентности к соответствующим информационным воздействиям.

6.2. Этапы информационных операций

Остановимся отдельно на этапности информационных операций. Очевидно, не существует единственного «стандартного» плана проведения как наступательных, так и оборонительных информационных операций. Можно лишь рассмотреть примерную, полученную путем обобщения некоторых уже реализованных информационных операций последовательность действий при их осуществлении.

На практике информационная операция как процесс информационного воздействия на массовое сознание, как правило, реализуется следующим образом: в результате предварительной разведки вырабатывается план следующего этапа — оперативного управления и намечаются соответствующие мероприятия оперативной разведки, которые являются

приближенной моделью решения, после чего реализуется оперативное управление противником. На этапе оперативной разведки определяется уровень отклонения первоначальной модели от реальности, и если оно незначительно, то реализуется первоначальный план. В противном случае строится новый план оперативного управления и управления противником. Далее цикл повторяется до тех пор, пока оперативная разведка не подтвердит используемую модель. При этом окончательное решение принимается с определенным оперативным риском.

Таким образом, процесс информационного воздействия охватывает такие основные этапы [Чхартишвили, 2004] (Рис. 89):

- предварительная разведка (preliminary intelligence, PI);
- выявление текущей обстановки, состояния противника (Op);
- управление противником (management of enemy, M) (информационное воздействие на противника с целью передачи ему сведений соответствующих замыслу управляющего);
- оперативная разведка (operational intelligence, OI) (проверка результатов рефлексивного управления);
- оперативное управление (operational management, OM) – действия управляющего для достижения требуемой цели.

При планировании или моделировании социальных процессов, в частности информационных операций, всегда необходимо учитывать, что общее поведение социальных систем невозможно определить, оперируя исключительно рафинированными математическими моделями. Это главным образом обусловлено тем, что такие процессы в большой степени зависят от социально-психологических факторов.

Различают два основных типа информационных операций — наступательные и оборонительные. Однако, на практике, большая часть информационных операций является смешанной. Кроме того, большинство процедур информационных операций относятся одновременно к наступательным и оборонительным. Каждый из типов информационных опера-

ций, включая приведенные выше основные этапы, подразумевает некоторые особенности и уточнения.

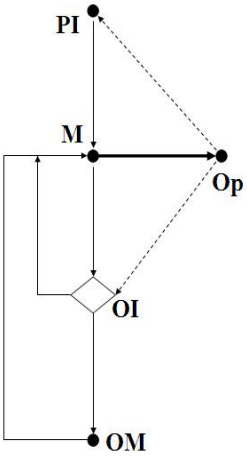


Рис. 89 – Основные этапы информационных операций

Особенностью наступательных информационных операций (информационных атак) является то, что объекты воздействия таких операций определены и планирование основывается на достаточно точной информации об этих объектах. Информационная атака чаще всего требует нахождения или создания информационного повода (для оборонительных информационных операций поводом может являться сама информационная атака противника), раскрутка этого повода, т.е. пропаганда (в отличие от мер контрпропаганды при оборонительных информационных операциях), а также необходимость принятия мер по препятствию информационному противодействию.

Таким образом, план типовой информационной операции включает совпадающие на верхнем уровне для информационных операций обеих типов такие этапы, как оценка, планирование, исполнение и завершающая фаза. Приведем более детальный перечень компонент информационных операций.

В наступательных информационных операциях можно выделить такие основные фазы:

1. Оценка необходимости проведения операции:
   1) определение цели, прогноз достижимости, степени влияния;
   2) сбор информации.
2. Планирование.
3. Исполнение информационного воздействия:
   1) нахождение или создание информационного повода;
   2) раскрутка информационного повода (пропаганда);
   3) оперативная разведка;
   4) оценка воздействия;
   5) препятствие информационному противодействию;
   6) корректировка информационного воздействия.
4. Завершающая фаза:
   1) анализ эффективности;
   2) использование позитивных результатов информационного воздействия;
   3) противодействие отрицательным результатам.

Типовая оборонительная информационная информация охватывает такие основные этапы:

1. Оценка:
   1) анализ возможных уязвимостей (целей);
   2) сбор информации о возможных операциях;
   3) определение возможных «заказчиков» информационных воздействий:
   - определение сфер общих интересов объекта и потенциальных «заказчиков»;
   - ранжирование потенциальных заказчиков по их интересам.
2. Планирование:
   1) стратегическое планирование оборонительной операции (явное или неявное):
   - определение критериев информационных воздействий;
   - моделирование информационных воздействий с учетом: связей объекта; динамики воздействия; «особых» (критичных) точек воздействия;
   - прогнозирование следующих шагов;

- расчет последствий.

2) тактическое планирование контроопераций.

3. Исполнение — отражение информационного воздействия:

1) выявление и «сглаживание» информационного повода;
2) контрпропаганда;
3) оперативная разведка;
4) оценка информационной среды;
5) корректировка информационного противодействия.

4. Завершающая фаза:

1) анализ эффективности;
2) использование позитивных результатов информационного воздействия;
3) противодействие отрицательным результатам.

Оперативное управление информационными операциями с использованием информационно-аналитических систем можно проиллюстрировать с помощью диаграммы, представленной на Рис. 90.

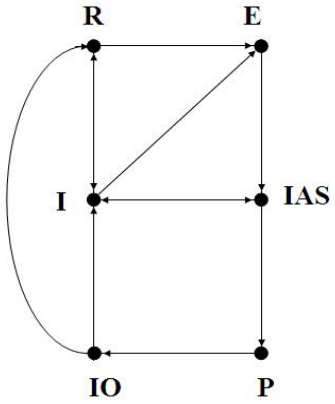


Рис. 90 – Диаграмма оперативного управления с использованием информационно-аналитических систем

В соответствии с приведенной диаграммой информация из реального мира (R) поступает в информационное пространство, в частности, в средства массовой информации (I)

либо непосредственно экспертам (E), также через средства массовой информации.

От экспертов или непосредственно из информационного пространства (например, с помощью средств контент-мониторинга) инфомация поступает в информационно-аналитическую систему (IAS). Информационно-аналитическая система передает лицам, принимающим решения (P), данные, которые определяют меры информационного воздействия на информационное пространство и непосредственно на объекты реального мира (людей, окружающую среду, компьютерные системы и т. д.).

6.3. Моделирование информационных операций

Моделирование можно рассматривать как один из способов решения проблем, возникающих в реальном мире, в частности, при планировании и проведении информационных операций. Чаще всего моделирование применяется в случаях, если эксперименты с реальными объектами невозможны, либо слишком затратные. Моделирование охватывает отображение реальной проблемы в мир абстракции, изучение, анализ и оптимизацию модели, и отображение оптимального решения обратно в реальный мир.

При моделировании существуют два альтернативных подхода — аналитическое и имитационное моделирование. Идеальные аналитические модели допускают строгое аналитическое решение или, по меньшей мере, постановку, например в виде систем дифференциальных уравнений. Однако, аналитические решения не всегда достижимы. Поэтому, особенно в последнее время, и особенно при решении задач из области социальной динамики все чаще применяются методы имитационного моделирования (*Simulation Modeling*). Имитационное моделирование представляет собой более мощное и практически незаменимое средство анализа социальных процедур. Имитационную модель можно рассматривать как множество правил, определяющих будущее состояние системы на основании текущего. При этом процесс моделирования заключается в наблюдении эволюции системы во времени по данным правилам, и, соответственно, оценки адекватности модели, когда это возможно.

Наиболее перспективным направлением моделирования информационных операций является математическое описание самоорганизации среды восприятия и распространения информации с учетом сложившихся в текущий момент условий. Самоорганизующиеся среды, для которых отсутствует центральный механизм управления, а развитие идет за счет множества локальных взаимодействий, изучаются теорией сложных систем. Эта теория охватывает такие отрасли знаний, как нелинейная физика, термодинамика неравновесных процессов, теория динамических систем. Взаимодействия между отдельными элементами сложных систем определяют возникновение сложного поведения при отсутствии централизованного управления. Для исследований подобного поведения применяются самые современные методы, которые охватываются междисциплинарной основой современной методологии — концепцией сложности. В настоящее время к теоретическим и технологическим основам этой концепции относятся теории детерминированного хаоса, фракталов и сложных сетей, синергетика, волновой (вейвлет) анализ, многоагентное моделирование, теория самоорганизованной критичности (изучающей динамическое развитие до критического состояния, характеризуемого сильными пространственно-временными флуктуациями, без внешнего управления [Bak, 1996]), теория перколяции (Percolation – протекание) и т.п.

Моделирование социальных процедур (информационные операции, безусловно, относятся к таковым) предполагает проведение вычислительных экспериментов, так как чаще всего возникают существенные ограничения, затрудняющие проведение «полевых» натуральных экспериментов.

При моделировании информационных операций вычислительный эксперимент позволяет сократить операции по уточнению ограничений, подбору исходных данных, выбору правил функционирования компонент модели и т.д. В этом случае появляется возможность учета случаев, трудно реализуемых на практике, используя реальные данные лишь для идентификации параметров математической модели. Вместе с тем математическое моделирование имеет свои ограничения, реальный мир оказывается сложным для моделирования с достаточным уровнем детализации и точности, т.е. более или менее достоверные математические модели настолько

сложны и многопараметричны, что не поддаются анализу и оценкам точными методами.

Отработать математические модели при планировании инфор-мационных операций можно лишь в процессе моделирования конкретных процедур, постоянно сопоставляя их с реальностью.

Выраженная цель методологии оценки информационных операций состоит в том, чтобы обеспечить своевременный и точный анализ возможных несоответствий между запланированной операцией и фактическим воздействием. Когда обнаруживаются существенные различия, которые влияют на вероятности успеха операции, аналитическая система должна сообщать об этом лицам, принимающим решения, для того, чтобы откорректировать текущие планы и решения. Вместе с тем, при планировании информационных операций нельзя действовать методом проб и ошибок, поэтому необходимо развивать методы, позволяющие обобщать ретроспективные данные, и на их основе проверять адекватность моделей.

В основу успешных моделей информационных операций закладываются синергетические подходы. Действительно, общество является сложной системой, каждая компонента которой характеризуется множеством признаков, имеет множество степеней свободы. При этом важным свойством этой системы является самоорганизация, которая является результатом взаимодействия таких компонент, как случайность, многократность, положительная и отрицательная обратная связь.

Особенностью математического моделирования информационных операций следует считать сравнительную простоту интерпретации получаемых результатов. Такие понятия, как «численность электората», «политический вес» и т.д., воспринимаются на интуитивном уровне даже без знакомства с точными, насколько они тут возможны, определениями. А это позволяет делать подобный анализ актуальных ситуаций предметом широкого обсуждения.

В силу того, что некоторые решения являются неустойчивыми по отношению к своим параметрам, значения таких параметров необходимо определять с высокой точностью. Для этого требуется комплекс методик, основанных не только на

обработке больших объемов статистических данных, но и на разносторонних социологических исследованиях.

В настоящее время реалистичной выглядит постановка задачи, состоящая в использовании математических моделей для прогнозирования возможных сценариев динамики социальных процессов на качественном уровне. В такой формулировке моделирование динамики занимает как бы промежуточный уровень между тем, что изложено здесь, и точным прогнозированием. И все же потребуется выбор значений параметров, которые бы в некотором разумном приближении соответствовали изучаемой ситуации, причем в большинстве случаев продуктивным оказывается использование относительных величин. Так, конечно, не получить достоверных данных о будущем развитии событий, но, скорее всего, можно составить более или менее адекватную картину того, что и как может произойти. А это уже не мало.

Для достижения успеха при этом отдельные информационные воздействия необходимо рассматривать как части единой информационной операции, точно так же, как артобстрел или авиационные атаки можно рассматривать как согласованные части военной операции.

При этом информационным операциям присущи такие основные особенности:

- информационные операции – это междисциплинарный набор методов и технологий в таких областях, как информатика, социология, психология, международные отношения, коммуникации, военная наука;
- до сих пор не существует стандартов проведения информационных операций;
- в развитии технологий информационных операций заинтересованы не только оборонные ведомства, но и многие правительственные и коммерческие организации;
- задача формирования научного подхода к информационным операциям является насущной и актуальной.

При проведении информационных операций существенно выявление содержания (знаний), вкладываемого в информацию, с учетом самых разнообразных аспектов – социаль-

ных, политических, религиозных, исторических, экономических, психологических, ментальных, культурных, присущих различным слоям общества. Поэтому в настоящее время имеет смысл рассматривать информационные операции шире, как операции, базирующиеся на знаниях (Knowledge Operations) [Burke, 2001].

Обычная сетевая информационная атака в веб-среде сегодня производится следующим образом: как правило, создается и некоторое время функционирует веб-сайт (назовем его «первоисточником»), при этом он публикует вполне корректную информацию. В час X на его странице появляется документ, обычно компромат на объект атаки, достоверный либо сфальсифицированный. Затем происходит так называемая «отмывка информации». Документ перепечатывают интернет-издания двух типов – заинтересованные в атаке и те, кому попросту не хватает информации для заполнения своего информационного поля. В случае претензий все перепечатывающие издания ссылаются на «первоисточник» и, в крайнем случае, по просьбе/требованию объекта атаки удаляют со своих веб-сайтов информацию. Первоисточник при необходимости также снимает информацию либо вовсе ликвидируется (после чего оказывается, что он зарегистрирован в Интернет на несуществующее лицо). Вместе с тем информация уже разошлась, задача первоисточника выполнена, атака стартовала.

Современное информационное пространство представляет собой уникальную возможность получения любой информации по выбранному вопросу при условии наличия соответствующего инструментария, применение которого позволяет анализировать взаимосвязь возможных событий или событий, которые уже происходят, с информационной активностью определенного круга источников информации. С другой стороны, при ретроспективном анализе любого процесса или явления интерес представляют определенные характеристики его развития, а именно:

- количественная динамика, присущая процессу или явлению, например, количество событий в единицу времени, или количество сообщений, имеющих отношение к нему;

- определение критических, пороговых точек, которые соответствуют количественной динамике явления;
- определение проявлений в критических точках, например, выявления основных сюжетов публикаций в СМИ относительно выбранного процесса или явления;
- после выявления основных проявлений явления в критических точках, эти проявления ранжируются, и исследуется динамика развития отдельных определенных проявлений до и после определенных критических точек;
- осуществляется статистический, корреляционный и фрактальный анализ общей динамики и динамики отдельных проявлений, на основе которых осуществляются попытки прогнозирования развития явления и отдельных его проявлений.

Для исследования взаимосвязи реальных событий и публикаций о них в сети Интернет авторами использовалась система InfoStream, обеспечивающая интеграцию и мониторинг сетевых информационных ресурсов.

Количество веб-публикаций в день по какой-либо теме, а особенно изменения (динамика) этой величины порой позволяют даже небольшим специалистам в предметной области делать более-менее точные выводы.

Получить данные подобной динамики можно, например, ежедневно заходя на сайты интеграторов новостей (news.yandex.ru, webground.su, uaport.net). Конечно, в лучшем положении пользователи профессиональных систем мониторинга типа Интегрум или InfoStream. Именно на основе последней системы получена удивительная статистика по количеству веб-публикаций по тематике эпидемий гриппа в разные периоды.

В качестве примера рассмотрим информационную кампанию, направленную против «Проминвестбанка», которая началась в конце сентября 2008 г.

С помощью системы контент-мониторинга InfoStream (www.infostream.ua) [Григорьев, 2007], сканирующей все основные информационные веб-сайты Украины в режиме реального времени, была определена динамика публикаций на

веб-сайтах сообщений, в которых упоминался «Проминвестбанк» за три месяца – сентябрь, октябрь и ноябрь (рис. 91). Эта динамика свидетельствует о небольшом количестве публикаций за первую половину сентября, однако затем пошел ряд публикаций, компрометирующих председателя правления В. Матвиенко, что вызвало относительно небольшой резонанс.

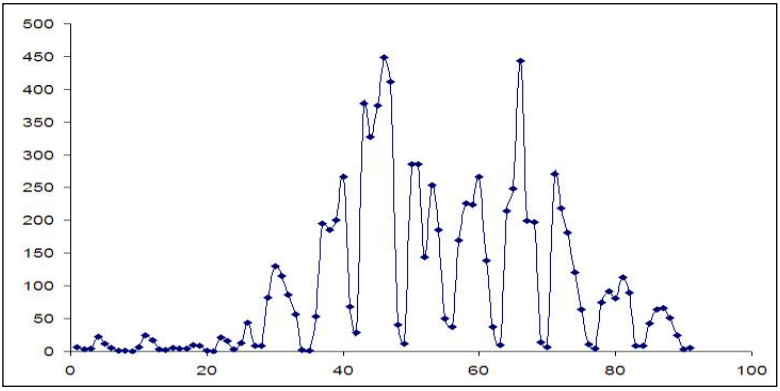


Рис. 91 – Динамика публикаций по теме «Проминвестбанк» за три месяца 2008 г.

Как оказалось впоследствии, эти публикации были лишь «артподготовкой». 26 сентября появились первые сообщения о возможном банкротстве банка (рис. 92), количество которых вполне соответствовало лавинообразному процессу, ограниченному лишь числом веб-сайтов, способных публиковать подобную информацию. Впрочем, этот процесс вышел на стабильно-средний уровень к декабрю 2008 г.

Нельзя утверждать, что лишь информационная атака через сеть Интернет привела банк к печальному состоянию, однако именно первые тревожные сообщения подорвали доверие многих вкладчиков, заставили их массово забирать свои сбережения из банка.

30 сентября появилось сообщение, что для спасения Проминвестбанка Национальный Банк Украины (НБУ) решил выделить ему 5 млрд. гривен рефинансирования, а 5 декабря появилось сообщение, что у «Проминвестбанка» появился новый владелец (рис. 93). После этого объемы публикаций о

«Проминвестбанке» существенно сократились, что свидетельствует не столько об его оздоровлении, сколько о системном кризисе банковской системы Украины, «уронившему» многие другие кредитные и банковские учреждения.

Kramatorsk.info **2008.09.26** 19:35
http://www.kramatorsk.info/?view&62181

В Донбассе вошла в активную фазу атака на Проминвестбанк. ПИБ заявляет, что это атака из-за рубежа

Сегодня в Донецкой области население организованно вышло к проходным **Проминвестбанка.**

Вести о том, что народные массы Донбасса штурмуют отделения ПИБа в Донецке, Авдеевке, Волновахе и пр. населенных пунктах Донецкой области, приходят в "Обком" с середины дня.

Никто из опрошенных нами экспертов не может пока сказать что-либо конкретное по данному поводу, кроме банальных констатаций: ПИБ - серьезный банк, он кредитует промышленный сектор Украины, Донецкое облотделение ПИБа - одно из крупнейших, борьба за него началась еще в середине 90-х годов... Ну а баннеры на киевских дорогах против нынешнего (неизменного) руководства ПИБа во главе с г-ном Матвиенко видели многие автомобилисты и пассажиры столичного транспорта.

"Обком" пока не готов сказать что-то определенное по поводу паники, которая охватила сегодня трудовой Донбасс - хотя сведения для определенных умозаключений, в принципе, имеются. Вместо этого мы предлагаем внимаю вкладчиков сообщение, поступившее от пресс-службы ПИБа:

"**Проминвестбанк** заявляет о стабильной работе, несмотря на дезинформацию в ряде СМИ о якобы приближающемся банкротстве банка.

Проминвестбанк, по оценкам зарубежных экспертов, стабильный банк и занимает в Украине второе место по надежности.

Массовая газетная атака на **Проминвестбанк** организована рейдерскими (бандитскими) группировками зарубежных агентов с участием высокопоставленных чиновников крупных государственных структур, которые по Конституции должны защищать отечественные предприятия и банки. Ложь, шантаж, направленные против банка, преследуют цель вынудить его к продаже иностранцам за комиссионное вознаграждение... Заявляем: банк не продается... **Проминвестбанк** останется украинским!", - говорится в сообщении.

Служба информационной поддержки **Проминвестбанка** также сообщает, что, несмотря на беспокойство вкладчиков, вызванное негативными публикациями о банке, все обязательства перед клиентами и вкладчиками выполняются, а структурные подразделения банка работают в нормальном режиме.

"Обком"

Рис. 92 – Одно из первых тревожных сообщений

Рис. 93 – Сообщения, завершившие экстремальную динамику интенсивности публикаций по теме «Проминвестбанк»

Буквально через неделю после описанных выше событий в Украине произошла еще одна публичная знаковая информационная атака, в этот раз на рынке страхования. Это была настоящая информационная операция против Национальной акционерной страховой компании (НАСК) «Оранта». В этом случае первоисточником компромата оказался не веб-сайт, а информационное сообщение, разосланное электронной почтой тысячам пользователей Интернет. В результате применения специальных технических приемов, оно разошлось с обозначением адреса пресс-службы объекта атаки. Итак, 10 декабря 2008 года в районе 11:30 в виде спама было разослано информационное сообщение, в котором говорилось о том, что

страховая компания «Оранта» заявляет о банкротстве. По предварительным данным, информация разлетелась по 1000 адресам, естественно, данные попали к конкурентам и в СМИ. В сообщении говорилось, что компания с 31 декабря 2008 года прекращает выполнять взятые перед клиентами обязательства.

В связи со случившимся НАСК «Оранта» обратилась в правоохранительные органы с просьбой расследовать данный инцидент и наказать виновных. Произошедшее с НАСК «Оранта» очень напоминало ситуацию с «Проминвестбанком», с этим согласились многочисленные эксперты. Ведь как банковский бизнес, так и страховой основываются на доверии клиентов, которое легче всего подрывается именно информационными атаками. По словам Олега Спилки, председателя наблюдательного совета НАСК «Оранта», «Это мероприятие готовилось целенаправленно для того, чтобы дискредитировать страховую компанию и подорвать ее репутацию». Не вдаваясь в детали возможных целей атаки (смена владельцев, борьба за блокирующий пакет акций, уничтожение компании и т.п.), с помощью ретроспективного анализа проследим за динамикой публикаций в сети Интернет, в которых упоминалась НАСК «Оранта».

На рис. 94 приведена посуточная динамика количества соответствующих публикаций. На этой диаграмме, кроме всего прочего, отчетливо виден спад интенсивности публикаций по данной теме в начале декабря 2008 г., что вполне можно воспринимать как некоторое «затишье перед бурей».

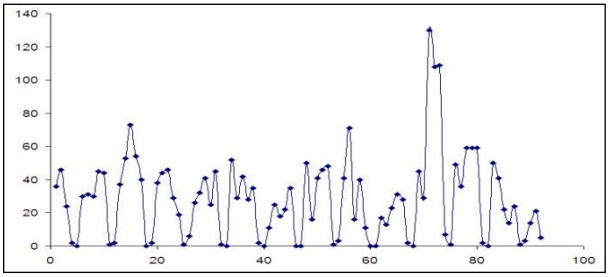


Рис. 94 – Интенсивность публикаций в Интернет по теме «Оранта»

Для анализа временных рядов в рамках исследования авторов применялся ΔL-метод. На рис. 95 представлена скейлограмма динамики рассматриваемого процесса с помощью метода (ΔL-метода) за второе полугодие 2008 года. Несмотря на отдельные пики в 16 и 55 день квартала, все же наибольший интерес представляет экстремум, приходящийся именно на 10–12 декабря.

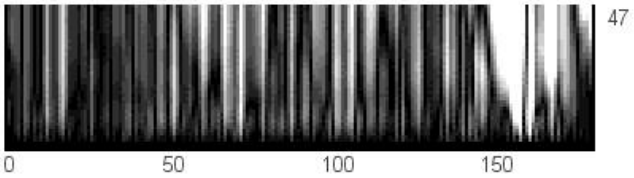


Рис. 95 – ΔL-диаграмма ряда публикаций по теме «Оранта»

Более детальная статистика публикаций по теме «Оранта» за декабрь 2008 года получена через интерфейс пользователя системы контент-мониторинга InfoStream (рис. 96).

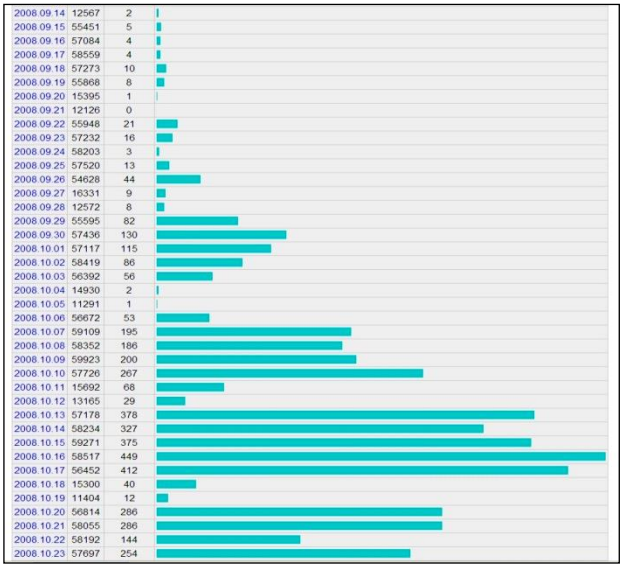


Рис. 96 – Детальная диаграмма интенсивности публикаций по теме «Оранта»

Проследим за ходом информационной операции, рассматривая сообщения, публикуемые в разные промежутки времени.

На Рис. 97 приведен список публикаций по теме «Оранта» в течение первых часов атаки. По словам Олега Спилки, в течение двух часов с начала атаки все почтовые серверы НАСК «Оранта» были выведены из строя, поэтому опровержение в сети задержалось.

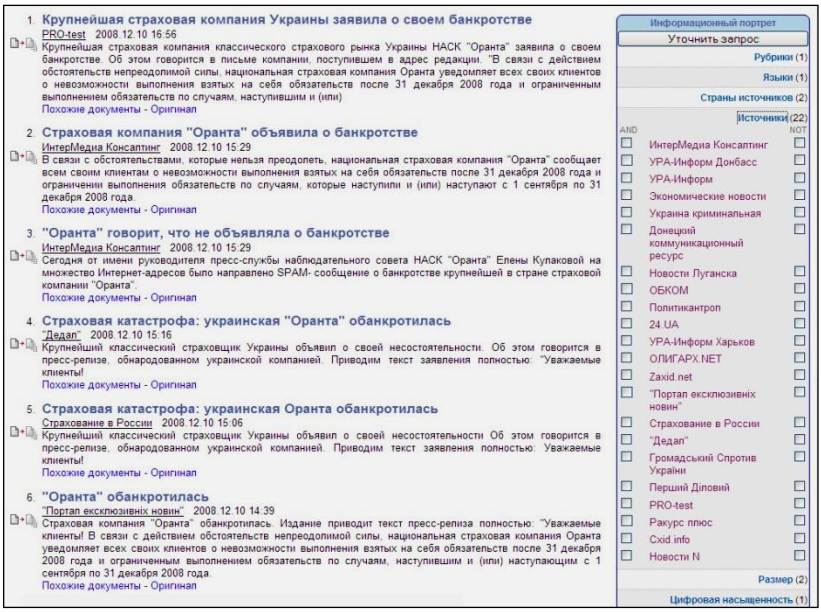


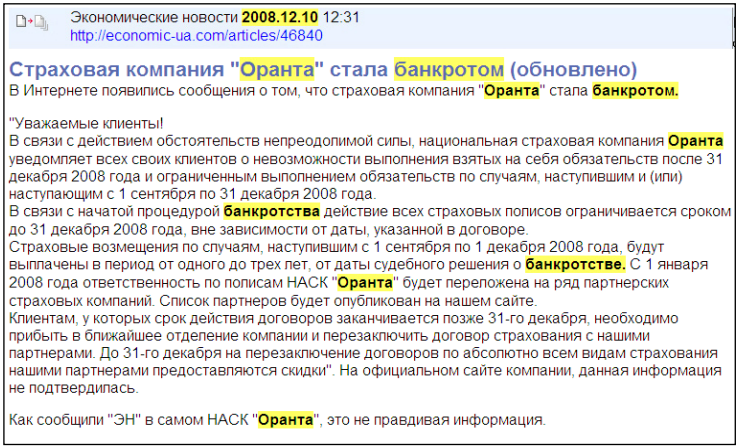
Рис. 97 – Первые часы атаки. Самые «оперативные» источники

В 12:31 на сайте «Экономические новости» появляется странное «обновленное» сообщение с парадоксальным последним предложением (рис. 98).

Далее руководство НАСК «Оранта» опубликовало в Интернете первые опровержения, не спеша обвинять конкурентов в происшедшем, а затем все же признав атаку целенаправленной и выгодной третьим лицам.

На рис. 99 приведен список публикаций, посвященных опровержению сообщения о банкротстве за следующий день

(11 декабря), а также наиболее активных источников, опубликовавших эти сообщения. Безусловный интерес аналитиков вызывает сравнение источников, приведенных на рис. 93 и 97.



Экономические новости **2008.12.10** 12:31
http://economic-ua.com/articles/46840

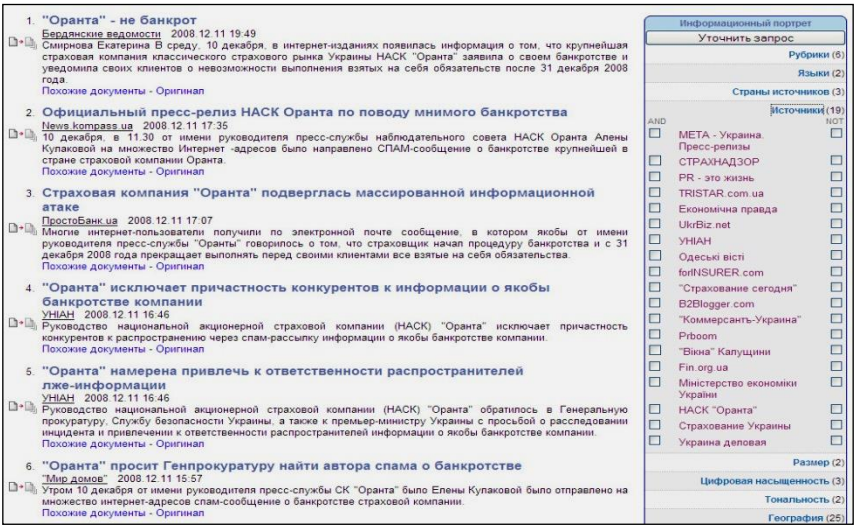
Страховая компания "Оранта" стала банкротом (обновлено)

В Интернете появились сообщения о том, что страховая компания "**Оранта**" стала **банкротом.**

"Уважаемые клиенты!
В связи с действием обстоятельств непреодолимой силы, национальная страховая компания **Оранта** уведомляет всех своих клиентов о невозможности выполнения взятых на себя обязательств после 31 декабря 2008 года и ограниченным выполнением обязательств по случаям, наступившим и (или) наступающим с 1 сентября по 31 декабря 2008 года.
В связи с начатой процедурой **банкротства** действие всех страховых полисов ограничивается сроком до 31 декабря 2008 года, вне зависимости от даты, указанной в договоре.
Страховые возмещения по случаям, наступившим с 1 сентября по 1 декабря 2008 года, будут выплачены в период от одного до трех лет, от даты судебного решения о **банкротстве.** С 1 января 2008 года ответственность по полисам НАСК "**Оранта**" будет переложена на ряд партнерских страховых компаний. Список партнеров будет опубликован на нашем сайте.
Клиентам, у которых срок действия договоров заканчивается позже 31-го декабря, необходимо прибыть в ближайшее отделение компании и перезаключить договор страхования с нашими партнерами. До 31-го декабря на перезаключение договоров по абсолютно всем видам страхования нашими партнерами предоставляются скидки". На официальном сайте компании, данная информация не подтвердилась.

Как сообщили "ЭН" в самом НАСК "**Оранта**", это не правдивая информация.

Рис. 98 – Опровержение?



1. **"Оранта" - не банкрот**
Бердянские ведомости 2008.12.11 19:49
Смирнова Екатерина В среду, 10 декабря, в интернет-изданиях появилась информация о том, что крупнейшая страховая компания классического страхового рынка Украины НАСК "Оранта" заявила о своем банкротстве и уведомила своих клиентов о невозможности выполнения взятых на себя обязательств после 31 декабря 2008 года.
Похожие документы - Оригинал

2. **Официальный пресс-релиз НАСК Оранта по поводу мнимого банкротства**
News.kompass.ua 2008.12.11 17:35
10 декабря, в 11.30 от имени руководителя пресс-службы наблюдательного совета НАСК Оранта Алены Купаковой на множество Интернет -адресов было направлено СПАМ-сообщение о банкротстве крупнейшей в стране страховой компании Оранта.
Похожие документы - Оригинал

3. **Страховая компания "Оранта" подверглась массированной информационной атаке**
ПростоБанк.ua 2008.12.11 17:07
Многие интернет-пользователи получили по электронной почте сообщение, в котором якобы от имени руководителя пресс-службы "Оранты" говорилось о том, что страховщик начал процедуру банкротства и с 31 декабря 2008 года прекращает выполнять перед своими клиентами все взятые на себя обязательства.
Похожие документы - Оригинал

4. **"Оранта" исключает причастность конкурентов к информации о якобы банкротстве компании**
УНІАН 2008.12.11 16:46
Руководство национальной акционерной страховой компании (НАСК) "Оранта" исключает причастность конкурентов к распространению через спам-рассылку информации о якобы банкротстве компании.
Похожие документы - Оригинал

5. **"Оранта" намерена привлечь к ответственности распространителей лже-информации**
УНІАН 2008.12.11 16:46
Руководство национальной акционерной страховой компании (НАСК) "Оранта" обратилось в Генеральную прокуратуру, Службу безопасности Украины, а также к премьер-министру Украины с просьбой о расследовании инцидента и привлечении к ответственности распространителей информации о якобы банкротстве компании.
Похожие документы - Оригинал

6. **"Оранта" просит Генпрокуратуру найти автора спама о банкротстве**
"Мир домов" 2008.12.11 15:57
Утром 10 декабря от имени руководителя пресс-службы СК "Оранта" было Елены Кулаковой было отправлено на множество интернет-адресов спам-сообщение о банкротстве страховой компании.
Похожие документы - Оригинал

Информационный портрет
Уточнить запрос
Рубрики (6)
Языки (2)
Страны источников (3)
Источники (19)
AND NOT
META - Украина. Пресс-релизы
СТРАХНАДЗОР
PR - это жизнь
TRISTAR.com.ua
Економічна правда
UkrBiz.net
УНІАН
Одеські вісті
forINSURER.com
"Страхование сегодня"
B2Blogger.com
"Коммерсантъ-Украина"
Prboom
"Вікна" Калущини
Fin.org.ua
Міністерство економіки України
НАСК "Оранта"
Страхование Украины
Украина деловая
Размер (2)
Цифровая насыщенность (3)
Тональность (2)
География (25)

Рис. 99. Сообщения с опровержением

Дальнейший спад публикаций по теме НАСК «Оранта» и возвращение его на нормальный «средний» уровень свидетельствует о том, что компания своими осторожными и точными действиями смогла с успехом противостоять информационной операции.

6.4. Выявление информационных операций

Для оперативного анализа информационной обстановки с целью выявления информационных операций применяются специализированные системы мониторинга информационного пространства (контент-мониторинга). Такие системы обеспечивают, во-первых, оперативность, которую не могут обеспечить традиционные поисковые системы (время индексации сетевого контента даже лучшими из них составляет от нескольких суток до нескольких недель). Во-вторых, полноту (как в плане источников, так и представления материалов источников), которую не всегда обеспечивают обычные агрегаторы новостей. И, в-третьих, необходимые аналитические средства, которые позволяют пользователю создавать аналитические отчеты, базирующиеся на публикациях по заданной тематике в необходимый период времени.

В плане профилактики информационных операций следует внимательно следить за динамикой публикаций о целевой компании, если есть возможность, с учетом тональности этих публикаций, пользоваться доступными аналитическими средствами, например, вейвлет-анализом. При этом следует ориентироваться на возможные модели информационных атак, например, если эта модель охватывает фазы: «фоновые публикации» — «затишье» — «артподготовка» — «затишье» — «атака» (рис. 52), то уже по первым трем компонентам можно с большой вероятностью предсказать грядущие события.

Приведенный выше план, очевидно, является идеальным, ориентированным исключительно на данные контент-мониторинга веб-ресурсов.

Конечно, в лучшем положении находятся пользователи профессиональных систем контент-мониторинга. Многие современные информационно-аналитические системы содержат в своем составе средства отображения статистики вхождения в базы данных понятий, соответствующих пользовательским запросам. В частности, авторами использовалась

подсистема статистики в рамках системы контент-мониторинга веб-пространства InfoStream, реализующая данную функциональность.

При изучении трендов информационных операций в качестве временных рядов рассматриваются именно ряды по количеству тематических публикаций за определенный промежуток времени (чаще всего – за сутки), соответствующие этим информационным операциям. Поэтому для выявления трендов исследуются информационные потоки, соответствующие тематикам информационных операций – тематические информационные потоки.

Приведенные в [Горбулін, 2009] тренды сообщений, соответствующие этапам информационной операции, приведены на Рис. 100. При этом аналитикам следует ориентироваться на такие модели, например, если мониторинг позволяет определить фазы: «фон» – «затишье» – «артподготовка» – «затишье» – «атака», то уже по первым трем компонентам можно с большой вероятностью предсказать будущие события.

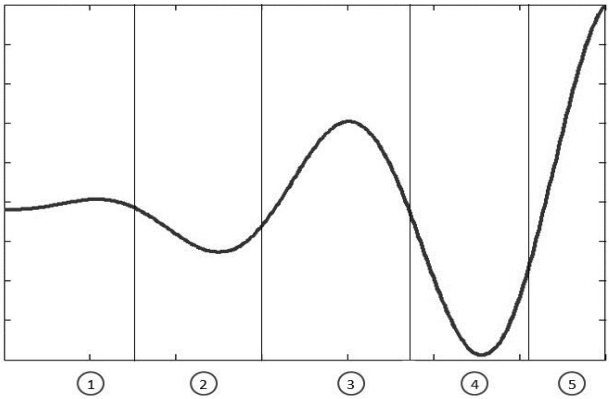


Рис. 100 – Динамика количества тематических сообщений во время проведения информационной операции: 1 – фон; 2 – затишье; 3 – «артподготовка»; 4 – затишье; 5 – атака/триггер роста

Следует отметить, что подобная динамика количества тематических сообщений при проведении информационных операций хорошо описывается известным уравнением распространения электромагнитных волн:

$$y = A + Bx\sin(x),$$

где x – время, A и B – константы, определяемые эмпирически.

Как известно, в настоящее время инновационная деятельность также косвенно измеряется количеством публикаций, относящимся к инновациям, существует несколько моделей инновационных процессов, среди которых можно выделить модель диффузии инноваций [Bhargava, 1993]. Вместе с тем, внедрение инноваций также можно считать информационными операциями. Поэтому обратимся к результатам соответствующих исследований. На рис. 101 приведена обоснованная в [Хорошевский, 2012] диаграмма количества публикаций, соответствующая тренду инновационной деятельности.

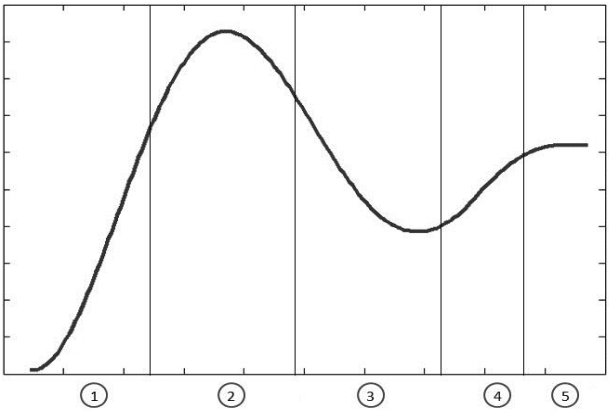


Рис. 101 – Диаграмма количества публикаций, соответствующих тренду инновационной деятельности: 1 – атака/триггер роста; 2 – пик завышенных ожиданий; 3 – утрата иллюзий; 4 – общественное осознание; 5 – продуктивность/фон

Объединяя графики, соответствующие началу информационной операции (Рис. 100) и тренду инновационной деятельности (Рис. 101), можно получить полный график, соответствующий отображению информационных операций в информационном пространстве (Рис. 102).

Предложенные модели полностью соответствуют реальным данным, которые экстрагируются системами контент-мониторинга [Додонов, 2009], [Ландэ, 2007]. Поэтому приведенные зависимости могут быть использованы как шаблоны для выявления информационных операций – как путем анализа ретроспективного фонда сетевых публикаций, так и для оперативного мониторинга появления некоторых их признаков в реальном времени. Как известно, для выявления информационных операций следует внимательно следить за динамикой публикаций по целевой теме и, если есть возможность, пользоваться доступными аналитическими средствами, средствами цифровой обработки данных и распознавания образов, например, вейвлет-анализом или полиномами Кунченко [Чертов, 2009].

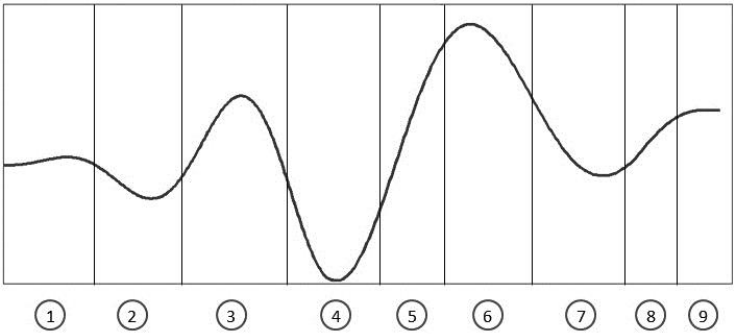


Рис. 102 – Обобщенная диаграмма, соответствующая всем этапам жизненного цикла информационных операций: 1 – фон; 2 – затишье;
3 – «артподготовка»; 4 – затишье; 5 – атака/триггер роста;
6 – пик завышенных ожиданий; 7 – утрата иллюзий;
8 – общественное осознание; 9 – продуктивность/фон

В качестве примера, на рис. 55 показана динамика публикаций в RUNet – тематических информационных потоков по запросам «Банки, Кипр», «Офшор», «Вирджинские острова» за март-апрель 2013 года, в период известных кризисных событий, полученная с помощью системы InfoStream. Как видно из Рис. 103, пик публикаций, связанных с банковским кризисом на Кипре приходится на 17-18 марта 2013 года, в то время, как большинство публикаций по Вирджинским ост-

ровам пришелся на 4–5 апреля, когда там, со значительно меньшими масштабами, стали проявляться события, подобные кипрским. При этом следует отметить слабую коррелированность динамики информационных потоков, связанных с Кипром и Вирджинскими островами. В этом случае коэффициент взаимной корреляции соответствующих числовых рядов составил всего 0,3. При этом отмечается высокий уровень взаимной корреляции рядов соответствующих тематикам «Офшор» и «Банки Кипра» (0,73), а также «Офшор» и «Вирджинские острова» (0,77).

По-видимому, проявления информационных операций в области офшорных банков в данном случае лучше всего увидеть при анализе более общей тематики – «Офшоры». На графике соответствующего числового ряда четко видны две области локальных экстремумов, соответствующих кризисным ситуациям на Кипре и на Вирджинских островах, а также фазы, соответствующие «затишьям» и «артподготовкам».

Можно высказать предположение, что если динамика частного информационного потока в какой-то момент начинает существенно отличаться от динамики потока, соответствующего более общей тематике (как в рассматриваемом случае, «Банки Кипра» и «Офшор»), то возможно проявление признаков начала информационной операции, относящейся к узкой тематике.

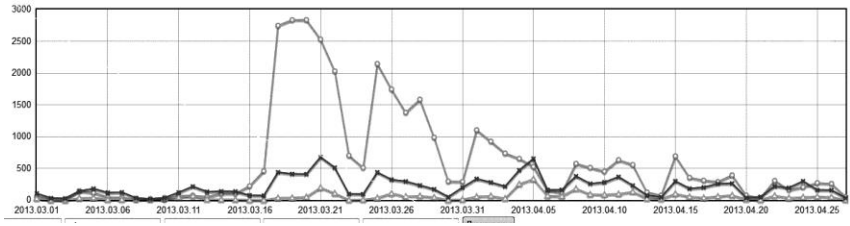
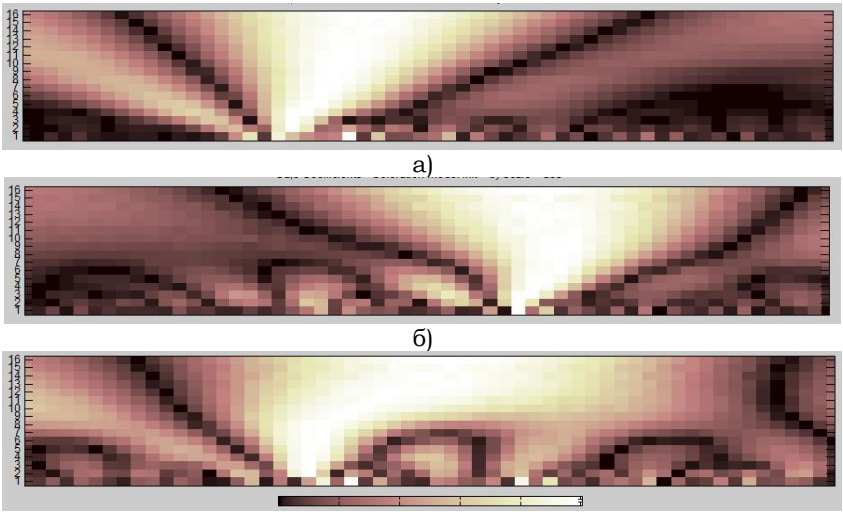


Рис. 103 – Диаграмма динамики тематических информационных потоков по запросам: o – «Банки Кипра»; Δ – «Вирджинские острова»; x – «Офшор»

При проведении вейвлет-анализа [Астафьева , 1996], [Buckheit, 1995] (рис. 56) было принято решение использова-

ния вейвлета «Мексиканская шляпа», как близкого по форме к диаграмме, приведенной на рис. 104.



а)

б)

в)

Рис. 104 – Вейвлет-спектрограммы, соответствующие динамике тематических информационных потоков по запросам: а – «Банки Кипра»; б – «Вирджинские острова»; в – «Офшор»

Рассматриваемые процессы четко просматриваются как на вейвлет-спектрограммах, так и на соответствующих им скелетонах (графиках линий экстремумов).

Приведенные модели и методы пригодны для описания общих тенденций динамики информационных процессов, однако, проблема прогнозирования остается открытой. По-видимому, более реалистичные модели могут быть получены с учетом дополнительного набора факторов, большинство которых не воспроизводятся во времени. Вместе с тем, структура правил, лежащих в основе функционирования большинства из доступных моделей, позволяет вносить соответствующие коррективы, например, искусственно моделировать случайные отклонения.

Отметим, что воспроизведение результатов во времени является серьезной проблемой при моделировании информа-

ционных процессов и составляет основу научной методологии. В настоящее время только ретроспективный анализ уже реализованных информационных операций остается относительно надежным способом их верификации.

Естественно, на практике ориентация лишь на единственный тип источников может привести к дефициту информации, необходимой для принятия решений, неточностям, а порой – к дезинформированности. Лишь применение комплексных систем, базирующихся на использовании многочисленных источников и баз данных, наряду с приведенными выше возможностями системы контент-мониторинга, может гарантировать эффективную информационную поддержку при противодействии информационным операциям.

Выделенные образцы поведения рядов интенсивностей тематических публикаций могут рассматриваться как шаблоны (образцы) функциональной зависимости. Эти шаблоны можно взять в качестве единого базисного элемента некоторого линейного пространства, т.е. в качестве порождающего элемента e для моделирования с помощью полиномов Кунченко [Чертов, 2009].

Тогда как линейную комбинацию линейно-независимых преобразований $f_1(e), f_2(e),..., f_n(e)$ соответствующего порождающего элемента можно построить полином P_n приближения n-го порядка к части выходного сигнала $f_s(e)$:

$$P_n = \sum_{\substack{k=0,\\ k\neq s}}^{n} c_k f_k(e),$$

где коэффициенты c_k определяются из условия обеспечения минимума расстояния между строящимся полиномом и сигналом. Элемент c_0 определяется выражением:

$$c_0 = \frac{\langle f_s(e), f_0(e)\rangle - \sum_{k=1,\ k\neq s}^{n} c_k \langle f_k(e), f_0(e)\rangle}{\langle f_0(e), f_0(e)\rangle},$$

а другие коэффициенты c_k – как решение системы линейных уравнений:

$$\sum_{k=1,\ k\neq s}^{n} c_k F_{i,k} = F_{i,s},\ i=1,...,n,\ i\neq s,$$

где центрированные корреляты $F_{i,k}$ также рассчитываются с помощью соответствующих преобразований:

$$F_{i,k} = \langle f_i(e), f_k(e)\rangle - \frac{\langle f_i(e), f_0(e)\rangle \cdot \langle f_k(e), f_0(e)\rangle}{\langle f_0(e), f_0(e)\rangle}.$$

Числовой характеристикой, которую можно использовать в критериях качества сопоставления сигнала с выделенным шаблоном, т. е. как меру приближения полинома Кунченко P_n к сигналу $f_s(e)$, можно считать коэффициент эффективности d_n:

$$d_n = \frac{\sum_{k=1,\ k\neq s}^{n} c_k \langle f_k(e), f_s(e)\rangle}{\langle f_s(e), f_s(e)\rangle}.$$

Рассмотренный метод распознавания определенных образцов с помощью построения пространства с порождающим элементом и поиска коэффициентов соответствующего полинома Кунченко может быть использован в любой проблемной области, в которой можно априори во временном ряду выделить определенные характерные шаблоны.

Таким образом, построив типовые модели поведения рядов интенсивности тематических публикаций во время проведения информационных операций и сопоставив шаблоны, полученные на их основе, можно использовать метод на основе полиномов Кунченко для определения (и предупреждения) возможной информационной атаки.

Динамика тематических информационных потоков определяется комплексом как внутренних, так и внешних нелинейных механизмов, которые должны быть отражены при моделировании (возможно, в неявном виде). Зачастую удовлетворительным оказывается упрощенное понимание тематического информационного потока как некоторой зависимой от времени величины, поведение которой описывается в

аналитическом виде нелинейными уравнениями. Сегодня при моделировании информационных потоков используются преимущественно аналитические нелинейные модели, применяются методы нелинейной динамики, теории клеточных автоматов, перколяции, самоорганизованной критичности [Ландэ, 2009], [Додонов, 2011].

Для анализа динамики реальных тематических информационных потоков (ТИП), и, соответственно, оценки их моделей необходимо каким-то образом получить соответствующую статистику, представленную в виде временных рядов.

Динамику реальных тематических информационных потоков (ТИП), например, отображает мультиагентная модель, в рамках которой отдельные документы ТИП ассоциируются с агентами, жизненный цикл которых – с жизненным циклом документов в информационном пространстве. Соответственно, все пространство мультиагентной модели ассоциируется с тематическим информационным потоком.

Предполагается, что в течение дискретных моментов времени происходит эволюция популяции агентов. При этом отдельные агенты могут:

1) «самозарождаться» (рождаться по причинам, возникающим вне рассматриваемого мультиагентного пространства);

2) «порождать» новых агентов;

3) «умирать» – исчезать из пространства агентов (соответст-вует утере актуальности документов);

4) получать ссылки от других агентов.

Каждый агент обладает «потенциалом», зависящим от его возраста (времени жизни на текущий момент – t), от авторитетности (ссылок, проставленных на него – ns) и плодовитости (количества порожденных непосредственно им агентов – k). Потенциал агента Pot определяется формулой:

$$Pot = \frac{1 + ns + k}{t}.$$

На рис. 105 приведен пример возможной динамики мультиагентой системы: процессы рождения новых агентов от существующих обозначены сплошными стрелками, процессы проставления ссылок на агентов представлены пунктирными стрелками, живые агенты – черными кругами, «мерт-

вые» агенты к моменту t = 5 – незаполненными окружностями.

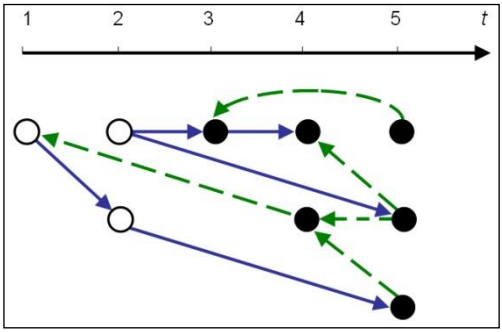


Рис. 105 – Фрагмент мультиагентного пространства

Итак, управляющие параметры модели следующие:

– вероятность «самозарождения» P_1;

– вероятность «рождения» от существующего: $P_2 \cdot Pot$;

– вероятность «смерти» агента: P_3 / Pot;

– вероятность ссылки на агента: $P_4 \cdot Pot$.

Варьирование этими четырьмя параметрами P_1, P_2, P_3 и P_4 позволили смоделировать типовые профили поведения ТИП.

На Рис. 106 представлены результаты численного моделирования количества агентов (ось ординат на графике) в рассматриваемой мультиагентной системе в зависимости от количества тактов модели (ось абсцисс).

Рассматриваемая модель эволюции пространства агентов при различных значениях управляющих параметров согласуется с динамикой реальных тематических информационных потоков, определенных с помощью системы InfoStream.

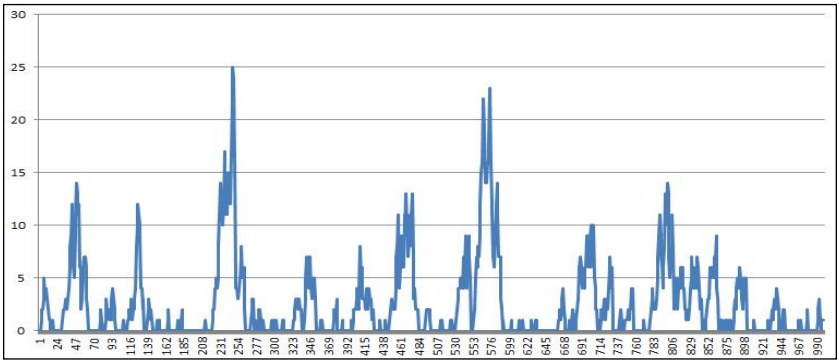


Рис. 106 – Динамика изменения количества агентов в модели

На практике ориентация лишь на единственный тип источников и математических моделей может привести к дефициту информации, необходимой для принятия решений, неточностям, а порой — к дезинформированности. Лишь применение комплексных систем, базирующихся на использовании многочисленных источников, баз данных, математических моделей, наряду с приведенными выше возможностями систем контент-мониторинга может гарантировать эффективную информационную поддержку при противодействии информационным операциям.

6.5. Пути противодействия информационным операциям

Рассмотренные практические примеры позволили выработать некоторую общую методику проведения оборонительной информационной операции с использованием системы контент-мониторинга веб-ресурсов. Допустим, объектом агрессивной информационной операции является компания «АБВ». Предлагается такие 12 шагов противодействия:

1) сбор информации с публикациями в «чужих» (не имеющих отношения к «АБВ», неаффилированных) СМИ о компании;

2) построение графика – динамики появления сообщений о компании «АБВ» в сетевых СМИ;

3) анализ динамики с ретроспективой в 6–12 месяцев с помощью методов анализа временных рядов. После этого анализируется контент публикаций в пороговых точках, определяются моменты, длительность, периодичность воздействия, привязка моментов воздействия к другим событиям из области интереса объекта;

4) определение источников, публикующих наибольшее количество негатива (публикаций с отрицательной тональностью) о компании «АБВ»;

5) определение «первоисточников» публикаций в СМИ – тех источников, которые первыми опубликовали негативную информацию;

6) определение вероятных «заказчиков» – владельцев или лиц, влияющих на издательскую политику отдельных СМИ;

7) определение сфер общих интересов компании «АБВ» и потенциальных «заказчиков» (путем выявления общих информационных характеристик – пересечений «информационных портретов» системы InfoStream, строящихся для объекта и «заказчика»), ранжирование потенциальных «заказчиков» по их интересам;

8) определение критериев информационных воздействий на основе самых рейтинговых интересов;

9) моделирование информационных воздействий, для чего находятся связи «заказчика» – наиболее связанные с ним персоны и организации, анализируется динамика воздействия со стороны заказчика и строится прогноз этой динамики, анализируется контент публикаций в пороговых точках кривой динамики – определяются критичные точки воздействия.;

10) прогнозируются дальнейшие шаги воздействия путем анализа аналогичной динамики публикаций для других компаний в ретроспективной базе данных системы InfoStream;

11) с учетом реалий и публикаций из ретроспективной базы данных оцениваются вероятные последствия;

12) организуется информационное (и не только) противодействие. Примеры публикаций в контексте противодействия находятся в ретроспективной базе данных.

6.6. Примеры информационных операций

Антимонопольная деятельность, создание в государстве конкурентной среды, предполагают борьбу с проявлениями монополизма на рынках товаров и услуг, в том числе, отражение соответствующих информационных операций, проводимых монополистами, проведение наступательных информационных операций.

Для осуществления антимонопольной деятельности со стороны государства, создания конкурентной среды необходимо использовать все доступные и легальные информационные и программные средства. Однако сегодня наблюдается реальный дефицит оперативной рыночной информации, определяемый как слабыми коммуникациями между отдельными органами власти, так и неполнотой, неточностью соответствующих официальных баз данных. С другой стороны, существует огромный информационный ресурс – веб-пространство.

Очевидно, что несмотря на такие преимущества, как оперативность и широкий охват информации, этот ресурс не может быть доказательным источником, однако, его нельзя отвергать в некоторых важных приложениях. Оперативность, свойственная веб-среде, в частности, имеет решающее значение при реализации концепции управления OODA, также известного как цикл Бойда. В переводе аббревиатура OODA означает «Наблюдение – Ориентация – Решение – Действие» [Ивлев, 2008]. Концепция OODA находит во всем мире широкое применение в управлении информационным противоборством, предотвращении информационным операциям. Очевидно, и в антимонопольной деятельности эта концепция может и должна найти применение путем реализации центров быстрого реагирования на монопольные проявления.

Общеизвестно, что антимонопольная деятельность – это комплекс мероприятий, направленных на ограничение деятельности монополий в рамках всего государства, а также создание соответствующего законодательства, в то время, как конкурентная разведка направлена на повышение конкурентоспособности лишь отдельных субъектов хозяйствования. Согласно этому частные задачи конкурентной разведки могут быть обобщены до уровня антимонопольной деятельности на уровне государства следующим образом:

1) сбор информации и своевременное информационное обеспечение соответствующих государственных органов;

2) выявление факторов риска, угроз конкурентной среды государства;

3) выявление факторов, влияющих на получение отдельными компаниями монопольных преимуществ;

4) выработка прогнозов и рекомендаций, влияющих на развитие конкурентной среды;

5) усиление благоприятных и локализация неблагоприятных факторов для развития конкурентной среды.

С помощью методов конкурентной разведки, которая становится современным направлением исследования поведения конкурентов на рынке, создаются альтернативные модели рынка для определения характеристик его участников и оптимизации тактики и стратегии развития субъектов хозяйствования на определенных рынках. Достижение таких целей требует использования эффективных приемов работы с информацией и ее элементами. Информация в этом смысле становится объектом в процессе исследования рынка и создания его модели.

Все приведенные задачи реализуются в рамках замкнутой схемы взаимодействия рыночной среды и виртуального информационного пространства.

Как известно, рыночная реальность находит свое отражение в виртуальном информационном пространстве, именно с ним работают эксперты-аналитики, которые готовят информацию, прогнозы для ЛПР, которые, в свою очередь, обеспечивают целенаправленное воздействие на рыночную среду.

По-видимому, все указанные функциональные компоненты конкурентной разведки могут использоваться и для общих задач, стоящих перед антимонопольными органами государства.

Возможности использования средств конкурентной разведки, в частности, средств контент-мониторинга, в антимонопольной деятельности проиллюстрируем на примере гречневого ценового коллапса в начале 2010 г. в Украине. Антимонопольный комитет Украины только в октябре 2011 г. (через полтора года!) обнаружил и наказал участников сговора на рынке гречки (Рис. 107), тогда как сотни пользователей

системы контент-мониторинга InfoStream могли видеть фигурантов дела уже в феврале 2010 г. в «информационном портрете» этой системы (Рис. 108).

Безусловно, система поддержки антимонопольной деятельности, как и системы конкурентной разведки, использующие Интернет как один из информационных ресурсов, должна настраиваться под специфику конкретных рынков. Она должна включать соответствующую классификацию, гибкие механизмы поиска, оперативной доставки данных, а также качественной оценки информации.

Одной из важнейших задач анализа информации при этом является определение ее достоверности, т.е. решение задачи анализа и фильтрации шума и ложной информации. После анализа достоверности информации должны следовать оценки ее точности и важности. Главным критерием достоверности данных на практике является подтверждение информации другими источниками, заслуживающими доверия.

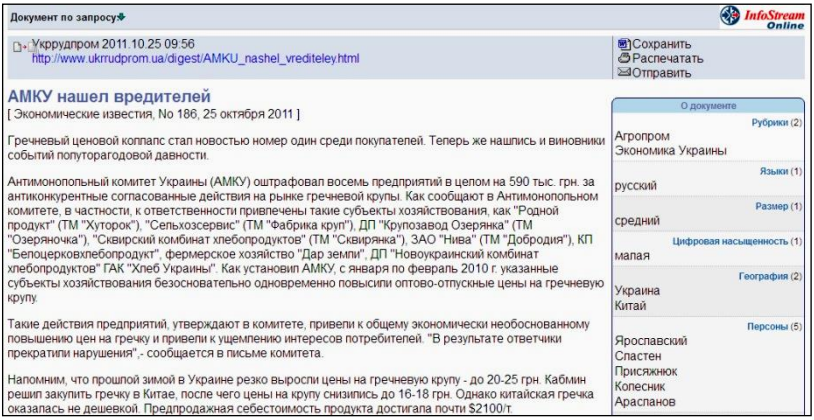


Рис. 107 - Виновники кризиса найдены (октябрь 2011 г.)

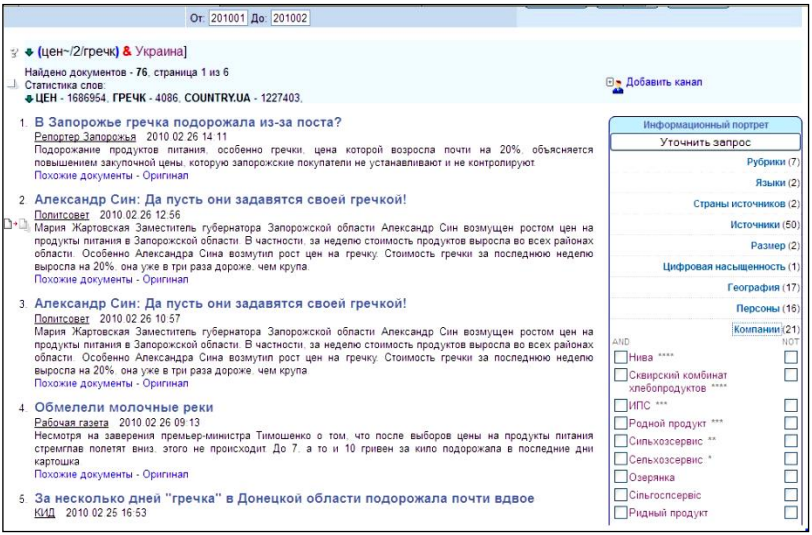


Рис. 108- Отражение «гречневого кризиса» 2010 г.

Условия исследования состояния рынка с помощью электронных средств, в частности, при проведении антимонопольной деятельности должны соответствовать современным условиям конкурентной разведки:

Во-первых, должны применяться методы и программные средства исследования информации полученной из открытых источников с соблюдением требований законодательства и этических норм.

Во-вторых, успех или неудача в решении практической задачи моделирования состояния рынка зависят от упрощения интегрированной информации, которую необходимо обработать.

В-третьих, достижение успеха при исследовании рынков связано с проблемой преодоления сложности доступа к информационным ресурсам из открытых источников, в том числе из сети Интернет.

Методология выявления антиконкурентных действий участников рынка по результатам анализа модели состояния рынка должна соответствовать возможностям имеющихся компьютерных средств и методов конкурентной разведки.

Например, данные, информацию и знания, получаемые в результате антимонопольных исследований, должны представляться в виде, соответствующем по структуре и форме разведывательной информации.

Современные средства, применяемые в конкурентной разведке в сетевой среде, обеспечивают:

– доступность необходимой части информации;

– огромный охват информации;

– оперативность, учет динамики информационных потоков.

В то же время, эти средства не могут заменить все инструменты, необходимые для антимонопольной деятельности. Для принятия решений в этой области требуется использование комплексных систем, которые позволяют добывать и обобщать информацию об объектах исследований из разных источников.

Таким образом, можно сделать вывод о том, что конкурентная разведка дополняет технологию поиска данных и информации в интернет-пространстве и целевое экстрагирование полезных понятий о состоянии и развитии товарного рынка с методами сбора, хранения, обработки и анализа данных, создает пространство интегрированной информации для анализа и формирования конкурентной политики.

Цели и средства антимонопольной деятельности обусловливают практические требования к созданию новых механизмов и технологий и требуют объединения различных по природе инструментов конкурентной разведки в соответствии с различными алгоритмами исследования.

Заключение

Актуальность конкурентной разведки в последнее время значительно возросла. Это связанно с такими процессами, как глобализация экономики, а, следовательно, и конкуренции, виртуализация экономики, развитие информационных технологий.

Широкому внедрению систем компьютерной конкурентной разведки способствуют и законодательные акты многих стран мира. Так, например, в США ещё в 1996 году был принят Закон о свободе информации, который обязал федеральные ведомства обеспечить гражданам свободный доступ ко всей своей информации. Ограничения касаются лишь материалов, имеющих отношение к национальной обороне, личных и финансовых документов, а также документов правоохранительных органов. Отказ в доступе к информации можно обжаловать в суде. Информация должна быть представлена в десятидневный срок, а споры разрешаться в течение 20 дней.

Во всем мире уже свыше 20 лет считают, что конкурентная разведка – это важнейшая функция современного менеджмента и главное условие динамичного и устойчивого развития бизнеса. Вместе с тем, как утверждает, гендиректор компании «Р-Техно» Роман Ромачев, «Если 10 лет назад конкурентные разведчики в первую очередь проверяли наличие у бизнес-партнеров криминальных связей, то сейчас они, как и на Западе, в большей степени добывают коммерческую информацию». Это подтверждают и данные исследования, проведенного Международным Центром конференций (МЦК) OnConference: большинство компаний используют конкурентную разведку для изучения состояния рынка (74 % респондентов) и конкурентов (64 %). Поиск, сбор и анализ информации помогает сформировать целостную картину конкурентной среды, установить причинно-следственные связи.

В настоящее время конкурентная разведка в сети Интернет обеспечивает доступность, огромный охват информации и высокую оперативность. Но она не может заменить другие виды и инструментальные средства бизнес-разведки. Для принятия серьезных решений необходимо использование комплексных систем, которые разрешают компоновать и

обобщать информацию об объекте исследований, полученную из разных источников с применением разных технологий.

Об актуальности конкурентной разведки на основе интернет-ресурсов говорят многочисленные публикации, тренинги, конференции. Сегодня задачи конкурентной разведки стимулируют развитие систем управления знаниями, глубинного анализа данных и текстов, с другой стороны наиболее развитые из этих систем в явном виде содержат аналитические блоки, специально ориентированные на задачи конкурентной разведки. Поэтому у пользователей имеется широкий выбор средств автоматизации аналитической деятельности. Причем уровни функциональности таких систем, может быть очень разнообразным – от простых информационно-поисковых программ, необходимых на этапе становления систем конкурентной разведки, до дорогих и ресурсоемких систем управления знаниями и глубинного анализа данных и текстов.

Для эффективного анализа современных информационных процессов на основе мониторинга информационных потоков из глобальных компьютерных сетей должны применяться современные методы, базирующиеся на нелинейном анализе, многие из которых нашли успешное применение в естественных науках. Современные подходы позволяют применять для анализа и моделирования даже общественных и информационных систем методы, апробированные в первую очередь в естественных науках. Анализ информационных потоков выступает фундаментом таких направлений как моделирование, проектирование и прогнозирование.

В то же время, приведенные модели и методы пригодны для описания общих тенденций динамики информационных процессов, однако, проблема прогнозирования остается открытой. По-видимому, более реалистичные модели могут быть получены с учетом дополнительного набора факторов, большинство которых не воспроизводятся во времени. Вместе с тем, структура правил, лежащих в основе функционирования большинства из доступных моделей, позволяет вносить соответствующие коррективы, например, искусственно моделировать случайные отклонения. Отметим, что воспроизведение результатов во времени является серьезной проблемой при моделировании информационных процессов, составляет ос-

нову научной методологии. В настоящее время только ретроспективный анализ уже реализованных информационных операций остается относительно надежным способом их верификации.

В настоящее время уже очевидно, что реальный прорыв в области интенсификации информационно-аналитической работы, как и в науке, возможен лишь в результате агрегирования различных направлений.

Краткий глоссарий

Автоматическое реферирование [Automatic text summarization] – автоматическое формирование краткого изложения исходного текстового материала либо путем выделения фрагментов информационного наполнения и последующего их соединения, либо методом генерации текста на основании выявления знаний из оригинала.

Авторское право [Copyright] – совокупность правовых норм, регулирующих отношения, возникающие в связи с созданием, использованием (изданием, исполнением, показом и т. д.) произведений науки, литературы или искусства – результатов творческой деятельности людей. Программы для компьютеров и базы данных также охраняются авторским правом.

Анализ социальных сетей [Social Networks Analysis, SNA] – методология анализа социальных сетей. Предметом анализа в SNA, в отличие от большинства традиционных социологических исследований, являются не атрибуты отдельных личностей, а структура их взаимосвязей в рамках того или иного сообщества (рабочей группы). В рамках анализа социальных сетей рассматриваются социальные отношения с точки зрения теории сетей, состоящих из узлов – личностей, участников сети и связей – отношений между ними.

Антимонопольная деятельность [Antitrust Activities] – комплекс мер, направленных на ограничение деятельности монополий, а также создание соответствующего законодательства.

База данных [Database] – совокупность данных, организованных по определенным правилам, предусматривающим общие принципы описания, хранения и манипулирования, независимая от прикладных программ. Является информационной моделью предметной области.

База данных аналитическая [Analytical Database] – база данных, которая содержит информацию, получаемую из других баз данных в форме итоговой информации, представляет наибольший интерес для пользователя или группы пользователей.

База данных полнотекстовая [Full-Text Database] – база данных, в которой хранятся записи полнотекстовых документов или их частей.

База данных фактографическая [Factographic Database] – база данных, содержащая фактографические данные – информацию, относящуюся непосредственно к предметной области.

Бенчмаркинг [Benchmarking] - инструмент анализа конкурента. Процесс определения, понимания и адаптации имеющихся примеров эффективного функционирования компании с целью получения информации, которая помогла бы предпринять шаги, направленные на улучшение деятельности компании. В равной степени включает в себя два процесса: оценивание и сопоставление. Одно из направлений стратегически ориентированных маркетинговых исследований.

Бизнес-процесс [Business Process] – система последовательных, целенаправленных и регламентированных видов деятельности, в которой посредством управляющего воздействия и при поддержке определенных ресурсов входы процесса преобразуются в выходы, представляющие ценность для потребителей.

Бизнес-разведка [Business Intelligence, BI] – 1) сбор и обработка данных из разных источников для выработки управленческих решений в целях повышения конкурентоспособности коммерческой организации; 2) структурное подразделение предприятия, выполняющее эти функции.

Блог [Blog, Web Log] – сетевой дневник одного или нескольких авторов, состоящий из записей в обратном хронологическом порядке. С помощью сервиса блогов можно создать свой онлайн-дневник, читать и комментировать дневники других пользователей, принимать участие в сообществах по определенным тематикам, создавать свои сообщества.

Блогосфера [Blogosphere] – совокупность (коллекция) всех блогов в сети Интернет; общее название для совокупности блогов.

Большие данные [Big Data] – серия подходов, инструментов и методов обработки данных больших объемов и значительного многообразия для получения человеко-читаемых результатов, эффективных в условиях их непрерывного прироста, распределения по многочисленным узлам вычислительной

сети. В качестве определяющих характеристик для больших данных отмечают «три V»: объем, в смысле величины физического объема, скорость в смыслах как скорости прироста, так и необходимости высокоскоростной обработки и получения результатов, многообразие, в смысле возможности одновременной обработки различных типов структурированных и полуструктурированных данных.

Веб-аналитика [Web Analytic] – измерение, сбор, *анализ*, представление и интерпретация информации о посетителях *веб–сайтов* в целях их улучшения и оптимизации. Основная задача веб-аналитики – *мониторинг* работы веб-сайтов, на основании которого определяется веб-аудитория и изучается поведение веб-посетителей для принятия решений по развитию и расширению функциональных возможностей веб-ресурса.

Веб-пространство [Web Space] - совокупность сайтов в Интернете (гипертекстовое пространство Интернета, www-пространство).

Веб-сайт [Website] – набор веб-страниц, составляющих единое целое (посвященных одной тематике либо принадлежащих одному и тому же автору), как правило, размещенных на одном и том же сервере, имеющих одно и то же доменное имя и связанных между собой перекрестными ссылками. Для прямого доступа клиентов к веб-сайтам на серверах разработан протокол HTTP.

Веб-форум [Web-Forum] – класс веб-приложений для организации общения посетителей, веб-сайт, предназначенный для проведения онлайн-дискуссий.

Виртуальная служба знакомств [Online Dating Service] – интернет-сервис, предоставляющий пользователям Интернета услуги по виртуальному общению с другими пользователями, аналог реальных служб знакомств.

Визуализация [Visualization] – комплекс методов представления результатов анализа данных в наиболее удобной для восприятия и интерпретации форме. Может использоваться для мониторинга процесса построения и работы различных аналитических моделей, проверки гипотез и других целей, связанных с проведением анализа.

Входная степень узла [In-Degree] – количество ребер графа, которые входят в узел.

Выходная степень узла [Out-Degree] – количество ребер графа, которые выходят из узла.

Геосоциальная сеть [Geosocial Networking] – вид социальных сетей, в которых используются геокодирование. Пользователи оставляют данные о своем местонахождении, что позволяет объединять и координировать их действия на основании того, какие люди присутствуют в тех или иных местах, или какие события происходят в этих местах.

Глубинный анализ данных [Data Mining] – технология анализа данных в базах или хранилищах данных, основанная на статистических методах и служащая для выявления заранее неизвестных закономерностей, а также для поддержки принятия стратегически важных решений.

Глубинный анализ текстов [Text Mining] – технология извлечения информации из текстовых данных на основе обнаружения в них закономерностей. Как правило, включает этапы структурирования исходного текста (обычно путем синтаксического анализа, добавления одних лингвистических структур и удаления других с последующей вставкой результатов в базу данных), поиска закономерностей в данных, оценивания и интерпретации результатов.

Глубинный веб [Invisible Web, Deep Web, Hidden Web] – часть веб-пространства, не индексируемую роботами поисковых систем. Информация, будучи недоступной для поиска, находится «в глубине» (англ. – Deep). Состоит из веб-страниц, динамически генерируемых по запросам к онлайн базам данных.

Граф связей [Communication Graph] – в социальных сетях – граф, предназначенный для идентификации связей между их участниками. С помощью графа можно визуализировать эти связи. Граф связей строится благодаря обмену контентами между людьми.

Дайджест [Digest] – информационный продукт (издание, статья, подборка), содержащий краткие аннотации и основные положения статей или в котором сжато передается содержание самых интересных публикаций за определенный период.

Дескриптор [(от лат. Descriptio – описание); Descriptor] – лексическая единица (слово, словосочетание, код) информационно-поискового языка, служащая для выражения основ-

ного смыслового содержания документов (текста). Используется для координатного индексирования документов и информационных запросов с целью последующего поиска.

Диаметр графа [Graph Diameter] – максимальное из расстояний между парами его вершин. Расстояние между вершинами определяется как наименьшее число ребер, которые необходимо пройти, чтобы добраться из одной вершины в другую.

Живучесть системы [System Survivability] – способность системы выполнять установленный минимальный объем своих функций при внешних воздействиях, не предусмотренных условиями нормальной эксплуатации, осуществлять выбор оптимального режима функционирования за счет собственных внутренних ресурсов, перестройки структуры, изменения функций отдельных подсистем и их поведения.

Извлечение знаний [Knowledge Extraction] – процесс получения из данных знаний в виде зависимостей, правил, моделей. Этапы: консолидация, очистка, трансформация, моделирование и интерпретация полученных результатов.

Извлечение (экстрагирование) информации [Information Extraction] – разновидность информационного поиска, при которой из электронных документов выделяется некая структурированная информация, т.е. катюгоризированные, семантически значимые данные по какой-либо проблеме или вопросу.

Извлечение фактов, понятий (Feature Extraction) – технология, обеспечивающая получение информации в структурированном виде. Включает три основных метода: Entity Extraction – извлечение слов или словосочетаний, важных для описания содержания текста; Feature Association Extraction – выявление связей между извлеченными понятиями; Event and Fact Extraction – извлечение сущностей, распознавание фактов и событий.

Имидж компании [Corporate Image] – устойчивые представление, которое компания создает о себе с помощью рекламы, формируя благоприятное представление у целевой аудитории. Это устойчивое представление потребителей, клиентов, партнеров и общественности о престиже компании, качестве её товаров и услуг, репутации руководителей.

Имитационное моделирование (Simulation Modeling) – метод исследования, при котором исследуемая система замещается моделью, с достаточной точностью описывающей реальную систему. Эту модель используют для экспериментов с целью получить информацию о реальной системе. Имитационное моделирование – частный случай математического моделирования. Существует класс объектов, для которых по разным причинам не разработаны аналитические модели или методы решения относительно полученной модели. В этом случае математическая модель замещается имитатором или имитационной моделью – логико-математическим описанием объекта.

Интернет [Internet] – глобальная информационная сеть, части которой логически связаны единым адресным пространством, основанным на стеке протоколов TCP/IP, их последующих расширений или других IP-совместимых протоколов. Обеспечивает, использует или делает доступным, публично или частным образом, коммуникационный сервис высокого уровня. Состоит из множества взаимосвязанных компьютерных сетей.

Интернет-разведка [Internet Intelligence] – сегмент конкурентной разведки, охватывающий процедуры сбора и обработки информации, проводимые с целью поддержки принятия управленческих решений, повышения конкурентоспособности коммерческих организаций исключительно из открытых источников из компьютерных сетей, большинство из которых являются надстроенными над сетью Интернет.

Интернет-чистильщики [Internet-Cleaners] – специалисты или службы, которые могут удалить из информационных ресурсов сети Интернет данные (как правило, негативную информацию о заказчике).

Информационная безопасность [Information Security] – состояние информации, информационных ресурсов и информационных систем, при котором с требуемой вероятностью обеспечивается защита информации (данных) от утечки, хищения, утраты, несанкционированного уничтожения, искажения, модификации (подделки), копирования, блокирования и т.п. Имеет три основные составляющие: конфиденциальность, целостность и доступность.

Информационно-аналитическая деятельность [Informational-Analytical Activity] – отрасль человеческой деятельности, призванная обеспечить информационные потребности общества с помощью аналитических и информационных технологий за счет обработки входной информации и получения качественно нового знания.

Информационно-аналитическая система, ИАС [Information-Analytical System, IAS] – класс информационных систем, предназначенных для аналитической обработки данных, а не для автоматизации повседневной деятельности организации. Объединяет, анализирует и хранит информацию, извлекаемую как из баз данных организации, так и из внешних источников. Входящие в состав ИАС хранилища данных обеспечивают преобразование больших объемов детализированных данных в обобщенную информацию, пригодную для принятия решений.

Информационно-поисковая система, ИПС [information retrieval system, IRS] – система, предназначенная для обеспечения поиска и отображения документов, представленных в базах данных. Ядро ИПС составляет поисковый механизм – программный модуль, который осуществляет поиск по запросу. ИПС, интегрированные с веб-технологиями, являются основой построения информационно-поисковых веб-серверов.

Информационное воздействие [Informational Influence] – возбуждение (торможение) в управляемой системе таких процессов, которые стимулируют желательный для управляющей стороны выбор. Этот способ воздействия на субъекта не предполагает, например, прямого выведения из строя части элементов его системы, но представляет собой передачу ему такой информации, которая натолкнет его на выбор определенного решения, при котором эти элементы потеряют свою эффективность.

Информационные операции [Information Operations] – информационное воздействие на массовое сознание (как на враждебное, так и на дружеское), воздействие на информацию, доступную конкуренту и необходимую ему для принятия решений, а также на информационно-аналитические системы (ИАС) конкурента, в том числе действия, направленные на физическое поражение ИАС, вывод из строя средств компьютерно-телекоммуникационной инфраструктуры.

Информационные ресурсы [Information Resources] – отдельные документы и отдельные массивы документов, документы и массивы документов в информационных системах, зафиксированные на соответствующих носителях информации, а также языковые средства, применяемые для описания конкретной предметной области и для доступа к данным и знаниям.

Информационные объекты, ИО [Informational Objects] – объекты, содержащие (несущие) информацию. Могут описываться непосредственно или в виде алгоритма их порождения.

Информационный портрет [Informational Portrait] – документ, который характеризует в компактной форме основное содержание текста - описанные в нем предметы, лица, ситуации и т.п.

Капча [Captcha] – нечеткое графическое изображение букв и цифр, которые требуется ввести с клавиатуры в определенное поле.

Классификация [Classification] – система распределения объектов по классам в соответствии с определенным признаком (основание классификации). Объекты необходимо классифицировать для выявления общих свойств информационного объекта, который определяется информационными параметрами (реквизиты). При классификации нужно соблюдать требования: полнота охвата; однозначность реквизитов; возможность включения новых объектов.

Кластерный анализ [Cluster Analysis] – многомерная статистическая процедура, выполняющая сбор данных, содержащих информацию о выборке объектов и затем упорядочивающая объекты в сравнительно однородные группы (кластеры).

Клики [Cliques] – подгруппы или кластеры, в которых узлы связаны между собой сильнее, чем с членами других клик.

Коммерческая тайна [Trade Secret] –сведения конфиденциального характера из любой сферы деятельности государственного или частного предприятия, разглашение которых может нанести материальный или моральный ущерб ее владельцам или пользователям (юридическим лицам).

Конкурентная разведка [Competitive Intelligence] – спланированные действия по систематическому сбору и ана-

лизу информации, проводимые с целью поддержки принятия управленческих решений, повышения конкурентоспособности коммерческих организаций.

Конкурентная среда [Competitive Environment] – результат и условия взаимодействия большого количества субъектов рынка. Образуется не только и не столько собственно субъектами рынка, взаимодействие которых вызывает соперничество, но в первую очередь – отношениями между ними.

Консалтинг [Consulting] – деятельность специализированных маркетинговых компаний, консультирующих производителей, продавцов, покупателей по вопросам в сфере экономики, управления, сбыта, ценообразования, продвижения продукции и др.

Консалтинговая компания (фирма) [Consulting Company (Firm)] – компания (фирма), выполняющая консультационные услуги по исследованию и прогнозированию рынков, разработке маркетинговых программ, поиску путей выхода из кризисных ситуаций; и др.

Консолидированная информация [Consolidated Information] – полученные из нескольких источников и интегрированные разнотипные информационные ресурсы (знания), которые в совокупности обладают признаками полноты, целостности, непротиворечивости и составляют адекватную информационную модель проблемной области с целью ее анализа обработки и использования в процессах поддержки принятия решений.

Контент [Content] – содержательное наполнение информационных ресурсов (напр., веб-сайтов) – тексты, графика, мультимедиа. Параметрами контента является его объем, актуальность и релевантность.

Контент-анализ [Content Analysis] – анализ содержания документов, который нацелен на измерение ряда качественных и количественных характеристик текста и на анализ зависимостей между ними.

Контент-мониторинг [Content Monitoring] – систематическое, непрерывное во времени сканирование и контент-анализ информационных ресурсов.

Конфиденциальная информация [Sensitive Information] – *информация*, которая представляет собой коммерческую или личную тайны и охраняется ее владельцем.

Конфиденциальность [Confidentiality] – свойство защищенности информации от несанкционированного доступа и попыток ее раскрытия пользователями, не имеющими соответствующих полномочий.

Коэффициент кластеризации [Clustering Coefficient] – величина, соответствующая уровню связности узлов в сети. Показывает, сколько ближайших соседей данного узла являются ближайшими соседями друг для друга, и равна отношению реального количества ребер, которые соединяют ближайших соседей данного узла, к максимально возможному.

Коэффициент посредничества [Betweenness] – параметр, показывающий, сколько кратчайших путей проходит через узел. Указывает на роль данного узла в установлении связей в сети.

Коэффициент центральности [Centrality] – параметр, который показывает «важность» или «влияние» определенного узла (кластера узлов) внутри графа (сети). Стандартные методы измерения «центральности» охватывают расчет центральность по посредничеству, центральность по близости, центральность собственного вектора, центральность по степени и др.

Лицо, принимающее решение, ЛПР [Decision Maker] – субъект (менеджер), наделённый определенными полномочиями и несущий ответственность за последствия принятого и реализованного управленческого решения. ЛПР – один или несколько человек (коллектив), на которых лежит ответственность за принятое решение.

Малый мир [Small Worlds] – один из видов графов, в котором большинство узлов не являются попарно соседними, но могут связываться друг с другом благодаря небольшому количеству переходов по ребрам графа. Граф (сеть) считается малым миром, если расстояние между двумя любыми случайно выбранными узлами в большинстве случаев не превышает двоичного логарифма от общего количества узлов.

Математическое моделирование [Mathematical Modeling] – процесс построения и изучения математических моделей – математических представлений о реальности.

Медиаактивность [Media Activity] – деятельность индивида по поиску, получению, потреблению, передаче, производству, распространению информации.

Метапоисковая система [Metasearch System] – поисковая система, не имеющая своего индекса, способная передавать запросы пользователя одновременно нескольким поисковым серверам, отбирать самые релевантные результаты, объединить их и представлять пользователю в виде документа со ссылками.

Мультиагентная система, МАС [Multi-Agent System, MAC] – система, образованная несколькими взаимодействующими интеллектуальными агентами. МАС могут быть использованы для решения таких проблем, которые сложно или невозможно решить с помощью монолитной системы.

Недобросовестная конкуренция [Unfair Competition] – нарушение общепринятых правил и норм конкуренции. Силовые и незаконные методы конкуренции (лишение конкурентов сырья, рынков сбыта, сбивание цен, промышленный шпионаж и т.п.).

Неструктурированная текстовая информация [Unstructured Text Information] – нестандартизированный и неформализованный текст, состоящий из предложений на естественном языке. Содержание текста – полнотекстовое изложение идей, смыслов и сюжетов (свободный текст).

Обнаружение знаний [Knowledge Discovery] – методика извлечения знаний (сведений) из источников информации.

Обнаружение знаний в тексте [Knowledge Discovery in Text] – процесс обнаружения новых, потенциально полезных и понятных шаблонов в неструктурированных текстовых данных.

Обнаружение знаний в базах данных [Knowledge Discovery in Databases, KDD] – процесс обнаружения полезных знаний в базах данных. Эти знания могут быть представлены в виде закономерностей, правил, прогнозов, связей между элементами данных и др. Главным инструментом KDD являются технологии Data Mining.

Общество аналитиков и профессионалов конкурентной разведки [Analysts Society and Competitive Intelligence Professionals] – областная общественная организация специалистов в области конкурентной разведки, созданная на базе кафедры Социальной информатики Харьковского национального университета радиоэлектроники в 2002 г.

Онтология [Ontology] – формализация некоторой области знаний с помощью концептуальной схемы, состоящей из структуры данных, их связи и правила, принятые в этой области. Сферы применения – моделирование бизнес-процессов, Семантический веб, искусственный интеллект.

Открытые источники [Open Sources] – информационные источники, легально распространяющие сведения, доступ к которым возможен на законных основаниях. Легальность и законность рассматривается только в контексте юрисдикции территории, на которой ведутся хозяйственные или иные операции.

Персональные данные [Personal Data] – любая информация, относящаяся к определенному или определяемому на основании такой информации физическому лицу, в том числе его фамилия, имя, отчество, год, месяц, дата и место рождения, адрес, семейное, социальное, имущественное положение, образование, профессия, доходы, другая информация. Персональные данные относятся к категории конфиденциальной информации; сведения или совокупность сведений о физическом лице, которое идентифицировано или может быть конкретно идентифицировано.

Пиринговая сеть [Peer-to-Peer, P2P] – компьютерная сеть, основанная на равноправии участников. В такой сети отсутствуют выделенные серверы, а каждый узел (peer) является как клиентом, так и сервером. В отличие от архитектуры клиент-сервера, позволяет сохранять работоспособность сети при любом количестве и любом сочетании доступных узлов.

Поисковая оптимизация [Search Engine Optimization, SEO] – комплекс мер для улучшения позиций веб-сайта в результатах выдачи сетевых поисковых систем по определенным запросам пользователей.

Построение семантической сети [Construction of Seman-tic Network] – одна из основных задач, решаемая в Text Mining – поиск ключевых понятий текста и установление взаимоотношений между ними, формирование структуры для представления знаний в виде ориентированного графа, в котором вершины – это понятия, а дуги – отношения. По такой сети можно осуществлять контекстную навигацию.

Промышленная контрразведка [Industrial Counter-Espionage] – деятельность по предупреждению промышленного шпионажа.

Промышленный шпионаж [Industrial Espionage] – форма недобросовестной конкуренции, при которой осуществляется незаконное получение, использование, разглашение информации, составляющей коммерческую, служебную или иную охраняемую законом тайну с целью получения преимуществ при осуществлении предпринимательской деятельности, а равно получения материальной выгоды.

Разведка [Exploration, Intelligance] – практика и теория сбора информации о противнике или конкуренте для обеспечения безопасности и получения преимуществ в области вооруженных сил, политики или экономики. Разведка может использовать как легальные методы сбора информации, так и нелегальные операции, попадающие под понятие «шпионаж».

Разведывательная деятельность [Intelligence Activities] – деятельность, включающая сбор информации, оценку ее достоверности и объединение отдельных фактов в общую картину.

Разведывательная информация [Intelligence Information] – осмысленные сведения, основанные на собранных, оцененных и истолкованных фактах, полученные в результате отбора, сопоставления, логической увязки и обобщения разведывательных данных и сведений в соответствии с заданием потребителя.

Разведывательный цикл [Intelligence Cycle] – в рамках конкурентной разведки – процессы, описывающие: целеуказание, сбор, обработку и анализ разведывательной информации, доведение целевой информации и выводов до заказчика.

Релевантность [Relevance] – мера соответствия получаемого результата желаемому. В информационном поиске – мера соответствия результатов поиска задаче поставленной в поисковом запросе.

Репутация [Reputation] – социальная оценка группы субъектов о человеке, группе людей или компании, сформировавшуюся на основе некоторых критериев.

Репутация компании [Company Reputation] – это комплекс оценочных представлений целевых аудитории о компа-

нии, сформированный на основе факторов репутации, имеющих значение для этой аудитории.

Ретроспективная информация [Retrospective Information] – сведения, содержащиеся в массивах данных, накопленных за значительный период времени, или полученные в результате поиска в этих массивах.

Ретроспективный анализ [Retrospective Analysis] – анализ, заключающийся в изучении тенденций, сложившихся за определенный период времени в прошлом.

Риск [Risk] – ситуативная характеристика деятельности, состоящая в неопределенности ее исхода и возможных неблагоприятных последствий в случае неуспеха.

Российское общество профессионалов конкурентной разведки, РОПКР [Russian Society of Competitive Intelligence Professionals, RSCIP] – зарегистрированное в г. Москве в 2002 г. юридическое лицо в виде некоммерческой организации в организационно-правовой форме некоммерческого партнерства. В состав основных целей РОПКР входит содействие продвижению, становлению, признанию, легитимности и развитию конкурентной разведки в России и странах СНГ, создание условий для объединения специалистов профессионально занимающихся конкурентной разведкой и специалистов других сфер деятельности, интересующихся теорией и практикой конкурентной разведки, создание условий для признания новой массовой профессии – специалист по конкурентной разведке.

Сбор данных [Data Collection] – процесс идентификации и получения данных от различных источников, группирования полученных данных и представление их в форме, необходимой для ввода в компьютер.

Семантическая сеть [Semantic Network] – способ представления знаний в виде ориентированного графа, в котором вершины соответствуют семантическим единицам языка (понятиям, объектам, действиям, ситуациям и т.п.), а ребра – свойствам или отношениям между ними.

Сетевая аналитика [Network Analytics] – совокупность средств и методов сбора из сетевой среды (в частности, из сети Интернет), преобразования, хранения, анализа, моделирования, доставки и трассировки данных, информации и

знаний при работе над задачами, связанными с принятием решений.

Сетевая мобилизация [Network Mobilization] – процесс объединения усилий участников социальных сетей для решения некоторых проблем, например, организации массовых выступлений, отражения агрессии, помощи пострадавшим и т.п. Возможности сетевой мобилизации зависят от структуры сети, ее топологии, параметров, динамики информации, циркулирующей в ней, возможности и вероятности восприятия информации узлами сети, возможности преобразования информации в узлах сети, возможности восстановления связей в сети после деструктивного воздействия на них.

Система конкурентной разведки [Competitive Intelligence Systems, CIS] – инфраструктура для осуществления конкурентной разведки; комплексная информационно-аналитическая система (ИАС) поддержки принятия решений в части анализа изменений условий ведения бизнеса и политической деятельности, на основе которого вырабатывается стратегия и тактика превентивных мероприятий, направленных на достижение конкурентных преимуществ и предотвращение влияния негативных факторов экономической и политической среды.

Слабые социальные связи [weakly social connections] – свойство социальных сетей, заключающееся в наличие связей (ребер с малыми весами) между удаленными в каком-то смысле узлами (например, отношения с далекими знакомыми и коллегами). Если эти связи проигнорировать, то сеть распадется на отдельные фрагменты. Слабые связи являются тем феноменом, который связывает социальную сеть в единое целое.

Слияния и поглощения [mergers and acquisitions, M&A] – класс экономических процессов укрупнения бизнеса и капитала, происходящих на макро- и микроэкономическом уровнях, в результате которых на рынке появляются более крупные компании взамен нескольких менее значительных.

Сообщество практиков конкурентной разведки, СПКР [Community Practice Competitive Investigation; CPCI] - сообщество практиков конкурентной разведки в России, существующее с 2004 г. Изначально СПКР было де-факто создано на Интернет-Форуме Бизнес-разведчиков, затем расши-

рялось, принимая в свои ряды новых специалистов из России, Украины, Беларуси. Одним из приоритетов в деятельности СПКР является активная пропаганда и продвижение конкурентной разведки в России и странах СНГ.

Социальная сеть [Social Network] – социальная структура, состоящая из узлов (которыми являются социальные объекты) и связей между ними. Объектами сетей могут быть предприятия, люди, Интернет-ресурсы и т.д. Существует множество социальных сетей, имеющих свою специфику, свойственные только им особенности, однако современные методы анализа данных применимы для любой из них вне зависимости от специфики.

Социальные медиа [Social Media] – совокупность онлайн-сервисов и интернет-приложений, которые позволяют пользователям общаться друг с другом в том числе, и в режиме реального времени. При этом пользователи могут обмениваться между собой мнениями, новостями, информацией, в том числе и мультимедийной. Социальные медиа базируются идеологической и технологической базе веб 2.0, позволяющих создание и обмен контентом, созданным самими пользователями (User-Generated Content).

Стратегическая деловая разведка [Strategic BI] - разведка, оказывающая помощь управлению в разработке их целостных планов и в проверке эффективности процесса видения. Охватывает сканирование окружения, анализ структуры отрасли, конкурентный анализ, анализ сценариев (планов действий), управление вопросами, технологическое прогнозирование, разработку типов конкурентных личностей и др.

Сценарий [Script, Scenario] – план выполнения процесса; определяет последовательность команд, которая указывает программе, как и в каком порядке, выполнять ту либо иную процедуру.

Сценарное планирование [Scenario Planning] – планирование вариантов развития событий (сценариев).

Тактическая деловая разведка [Tactical BI] – разведка, помогающая компании в ее повседневной работе, используемая сотрудниками непосредственно на своих участках на уровне ежедневного контроля. Охватывает анализ нужд по-

купателя, цены конкурента, анализ продукции и услуг конкурента, производства конкурента и др.

Теория сложных сетей [complex networks] – междисциплинарная область знаний, возникшая на базе эмпирических исследований реальных сетей, прежде всего, компьютерных и социальных. В рамках этой теории сложная сеть представляет собой граф (сеть) с нетривиальными топологическими особенностями, которые не встречаются в простых сетях, таких как решетки или случайные графы, но часто встречаются в реальности. Теория сложных сетей изучает характеристики сложных сетей, учитывая не только топологию сетей, но и статистические феномены, распределение весов отдельных вершин и ребер, эффекты протекания и проводимости в сетях и т.п.

Управление знаниями [Knowledge Management] – процессы, благодаря которым создаются, сохраняются, распределяются и применяются основные элементы интеллектуального капитала, необходимые для успеха организации. Существует пять основных технологий, которые поддерживает управление знаниями: бизнес-разведка (Business Intelligence), сотрудничество (Collaboration), трансфер знаний (Knowledge Transfer), обнаружение знаний (Knowledge Discovery) и определение экспертов (Expertise Location).

Управление репутацией [Reputation Management] – методы мониторинга репутации персоны или компании, выявления фактов которые вредят ей, и использование каналов обратной связи с потребителем для реакции или раннего выявления возможных негативных последствий для репутации.

Управление репутацией в Интернете [Online Reputation Management, ORM] – один из современных способов манипулирования интернет-контентом (популяризация информативных площадок, написание пресс-релизов, статей и отзывов) с целью создания положительного или отрицательного образа компании или персоны в интернете. Для осуществления ORM требуются специалисты: копирайтеры, редакторы, SEO-специалисты, дизайнеры и программисты. Цель специалистов ORM – получить привлекательный образ компании или персоны, повысить прибыльность и эффективность предприятия.

Управление рисками [Risk Management] – процесс принятия и выполнения управленческих решений, направленных

на снижение вероятности возникновения неблагоприятного результата и минимизацию возможных потерь, вызванных его реализацией. Цель риск-менеджмента в сфере экономики – повышение конкурентоспособности хозяйствующих субъектов с помощью защиты от реализации рисков.

Управленческое решение [Management Decision] – директивный акт целенаправленного воздействия на объект управления, основанный на анализе данных, характеризующих конкретную управленческую ситуацию, определение цели действий, и содержащий программу достижения цели.

Фактографическая база данных [Factographic Database] – база данных, содержащая фактографические данные – информацию, относящуюся непосредственно к предметной области.

Фактографическая информация [Factographic Information] – описание фактов, сгруппированных по определенным системообразующим признакам.

Фотохостинг [Photo Hosting] – веб-сайт, позволяющий публиковать изображения (например, цифровые фотографии) в сети Интернет. Фотохостинг может использоваться для размещения, хранения и показа изображений другим пользователям сети. Основное преимущество, которое предоставляет фотохостинг пользователям – удобство демонстрации фотографий. Автор может легко поделиться гиперссылкой, ведущей на фотографию, с любым человеком, имеющим доступ к сети Интернет.

Цифровая тень [Digital Shadow] – информация о пользователе, создаваемая без его участия, которая возникает и накапливается, когда кто-то ищет пользователя через поисковые системы, происходит электронная почтовая рассылка по спискам, в которых он фигурирует и во многих других случаях. Кроме «цифровых теней открытого доступа», создаются и копятся «цифровые тени ограниченного доступа» – записи камер наблюдения, банковские транзакции, биллинги интернет-магазинов, сервисов продажи билетов, телефонных звонков и др.

Цифровой след [Digital Footprint] – информация, которая оставляется самим пользователем при работе в Сети и по которой можно не только его идентифицировать, но и «привя-

зать» к определенным действиям, событиям, восстановить какие-то фрагменты биографии.

Экспертная оценка [Expert Estimates] – основанная на суждениях специалистов количественная или качественная оценка процессов или явлений, не поддающихся непосредственному измерению.

Экстрактор [Extractor] - программа, собирающая данные из исходных систем (выделение в тексте сложных элементов и специальных конструкций, отличающихся особого вида написанием, – наименований юридических лиц, товаров, адресов, номеров и т.п.).

Эксцентричность узла [Eccentricity] – наибольшее из геодезических расстояний (минимальных расстояние между узлами) от данного узла сети к другим.

Литература

[Астафьева, 1996] Астафьева Н.М. Вейвлет-анализ: основы теории и примеры применения // Успехи физических наук, 1996. – 166. – No 11. – P. 1145-1170.

[Берд, 2007] Берд К. Модель OSINT // Компьютерра, 2007. – № 22.

[Горбулін, 2009] Горбулін В.П., Додонов О.Г., Ланде Д.В. Інформаційні операції та безпека суспільства: загрози, протидія, моделювання: монографія. – К.: Інтертехнологія, 2009. – 164 с.

[Григорьев, 2007] Григорьев А.Н., Ландэ Д.В., Бороденков С.А., Мазуркевич Р.В., Пацьора В.Н. InfoStream. Мониторинг новостей из Интернет: технология, система, сервис: научно-методическое пособие. – Киев: Старт-98, 2007. – 40 с.

[Губанов, 2009] Губанов Д.А., Новиков Д.А., Чхартишвили А.Г. Модели репутации и информационного управления в социальных сетях // Математическая теория игр и ее приложения, 2009. – № 2. – С. 14-37.

[Джилад, 2010] Джилад Б. Конкурентная разведка. Как распознавать внешние риски и управлять ситуацией – СПб.: Питер, 2010. – 320 с.

[Додонов, 2009] Додонов О.Г., Ланде Д.В., Путятін В.Г. Інформаційні потоки в глобальних комп'ютерних мережах. – К: Наук. думка, 2009. – 295 с.

[Додонов, 2010] Додонов А.Г., Ландэ Д.В. Живучесть информационных сюжетов // Материалы XI Международной научно-практической конференции «Информационная безопасность». – Ч. 2. – Таганрог: Изд-во ТТИ ЮФИ, 2010. – С. 179-183.

[Додонов, 2011] Додонов А.Г., Ландэ Д.В. Живучесть информационных систем. – К.: Наук. думка, 2011. – 256 с.

[Додонов, 2013] Додонов А.Г., Ландэ Д.В., Коженевский С.Р., Путятин В.Г. Компьютерные информационно-аналитические системы и хранилища данных. Толковый словарь. – К.: Феникс; ИПРИ НАН Украины, 2013. – 554 с.

[Доронин, 2011] Доронин А. Бизнес-разведка. – М.: Ось-89, 2003. – 704 с.

[Дудихин, 2004] Дудихин В.В., Дудихина О.В. Конкурентная разведка в Интернет. – М.: АСТ, НТ Пресс, 2004. – 240 с.

[Ермаков, 2005] Ермаков Н.С., Иващенко А.А., Новиков Д.А. Модели репутации и норм деятельности. М.: ИПУ РАН, 2005. – 67 с.

[Іващенко, 2006] Основи методики розслідування незаконного збирання та розголошення комерційної таємниці // Юридичний журнал, 2006. – № 8. – С. 48-66.

[Ивлев, 2008] Ивлев А.А. Основы теории Бойда. Направления развития, применения и реализации (монография). – М., 2008. – 64 с.

[Калиновский, 2012] Калиновский Я.А., Бояринова Ю.Е. Высокоразмерные изоморфные гиперкомплексные числовые системы и их использование для повышения эффективности вычислений. –К.: Инфодрук, 2012. – 183 с.

[Киселев, 2005] Киселев С. Модель информационной системы бизнес-разведки // Открытые системы, 2005. – № 5-6. – С. 60-66.

[Ковальчук, 2012] Ковальчук А. Практика и секреты заработка в Интернете. Управление репутацией // Выпуск 30, 2012 (on-line: http://www.trustlink.ru/subscribe/ show/35)

[Кондратьев, 2010] Кондратьев А. Разведка с использованием открытых источников информации в США // Зарубежное военное обозрение, 2010. – №9. – С. 28-32.

[Кононов, 2003] Кононов Д.А., Кульба В.В., Шубин А.Н. Базисные понятия моделирования информационного управления в социальных системах // Труды международной научно-практической конференции «Теория активных систем». – М.: Институт проблем управления им. В.А. Трапезникова РАН, 2003. –Т 2. – С. 125-129.

[Кочергов, 2009] Кочергов Д. Один шаг, который может стать последним // Экономика бизнеса, 2009. – № 13 (9279).

[Кузнецов, 2006] Кузнецов С.В. Как вести бизнес-разведку в «невидимом» интернете? // «CNews», 07.09.06.

[Кульба, 1999]: Кульба В.В., Малюгин В.Д., Шубин А.Н., Вус М.А. Введение в информационное управление. Учебно-методичес-кое издание. – СПб.: Изд-во С.-Петербургского ун-та. 1999. – 116 с.

[Кульба, 2004] Кульба В.В., Кононов Д.А., Косяченко С.А., Шубин А.Н. Методы формирования сценариев развития социально-экономических систем. – М.: СИНТЕГ, 2004. – 296 с.

[Ландэ, 2005] Ландэ Д.В. Поиск знаний в Internet. Профессиональная работа. – М.: Диалектика, 2005. – 272 с.

[Ландэ, 2007] Ландэ Д.В., Снарский А.А., Брайчевский С.М., Дармохвал А.Т. Моделирование динамики новостных текстовых потоков // Интернет-математика 2007: Сборник работ участников конкурса. – Екатеринбург: Изд-во Урал. ун-та, 2007. – С. 98-107.

[Ландэ, 2009] Ландэ Д.В., Снарский А.А., Безсуднов И.В. Интернетика: Навигация в сложных сетях: модели и алгоритмы. – М.: Либроком (Editorial URSS), 2009. – 264 с.

[Ландэ, 2010] Ландэ Д.В. Глубинный web – информационная среда для бизнес-аналитика // Информационные технологии для менеджмента, 2010. – № 9. – С. 28-32.

[Ландэ, 2013] Ландэ Д.В. Метод визуализации зон нестабильности в рядах измерений // Информационные технологии и безопасность. Оценка состояния: Материалы международной научной конференции ИТБ-2013. – К.: ИПРИ НАН Украины, 2013. – С. 105-113.

[Ландэ, Брайчевский, 2010] Ландэ Д.В., Брайчевский С.М, Дармохвал А.Т, Жигало В.В. Архитектура системы охвата информационных связей объектов мониторинга // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог». – Вып. 9 (16). – М.: Изд-во РГГУ, 2010. – С. 272-278.

[Ландэ, Прищепа, 2007] Ландэ Д., Прищепа В. Школа вебразведки. Инструменты и источники // Телеком, 2007. – № 7-8. – С. 46-49.

[Ланде, Фурашев, 2012] Ланде Д.В., Фурашев В.М. Основи інформаційного і соціально-правового моделювання: монографія. – К.: ПанТот, 2012. – 144 с.

[Нежданов, 2010] Нежданов И. Технологии разведки для бизнеса. – М.: Ось-89, 2009. – 400 с.

[Новиков, 2002] Новиков Д.А., Чхартишвили А.Г. Теория управления организационными системами – М.: Синтег, 2002. – 227 с.

[Новиков, 2007] Новиков Д.А. Теория управления организационными системами. 2-е изд. – М.: Физмалит, 2007. – 584 с.

[Печенкин, 2004] Печенкин И.А. Информационные технологии на службе разведки // Конфидент, 2004. – № 4. – С. 28-41.

[Прескотт, 2003] Прескотт Джон Е., Миллер Стивен Х. Конкурентная разведка: Уроки из окопов. –М.: Альпина Паблишер, 2003. – 336 с.

[Расторгуев, 2006] Расторгуев С.П. Информационная война. Проблемы и модели. Экзистенциальная математика. – М.: Гелиос АРВ, 2006. – 240 с.

[Робинсон, 2016] Робинсон Ян, Вебер Джим, Эифрем Эмиль. Графовые базы данных: новые возможности для работы со связанными данными. – М.: ДМК Пресс, 2016. – 256 с.

[Хан, 2000] Хан У., Мани И. Системы автоматического реферирования // Открытые системы, 2000. – № 12.

[Хорошевский, 2013] Хорошевский В.Ф. Семантические технологи: ожидания и тренды // Открытые Семантические технологии проектирования интеллектуальных систем – Open Semantic Technologies for Intelligent Systems (OSTIS-2012): материалы II Междунар. научн.-техн. конф. (Минск, 16-18 февраля 2012 г.). – Минск: БГУИР, 2012. – С. 143-158.

[Черных, 2013] Черных Е. Нам от АНБ не спрятаться, не скрыться // Комсомольская правда, 24 июня 2013.

[Чертов, 2009] Четов О.Р. Поліноми Кунченка для розпізнавання образів // Вісник НТУУ «КПІ» Інформатика, управління та обчислювальна техніка, 2009. – № 50. – С. 105-110.

[Чхартишвили, 2004] Чхартишвили А.Г. Теоретико-игровые модели информационного управления. М.: ЗАО «ПМСОФТ», 2004. – 227 с.

[Bak, 1996] Bak P. How nature works: The science of self-organized criticality. – New York: Springer-Verlag Inc., 1996. – 212 p.

[Bhargava, 1993] Bhargava S.C., Kumar A., Mukherjee A. A stochastic cellular automata model of innovation diffusion // Technological forecasting and social change, 1993. – 44. – № 1. – P. 87-97.

[Bjorneborn, 2004] Bjorneborn L., Ingwersen P. Toward a basic framework for webometrics. Journal of the American Society for Information Science and Technology, 2004. –55(14): 1216-1227.

[Boyd, 2012] Boyd, D.; Crawford, K. (2012). "Critical Questions for Big Data". Information, Communication & Society. 15 (5): 662–679. doi:10.1080/1369118X.2012.67887

[Buckheit, 1995] Buckheit J., Donoho D.Wavelab and reproducible research // Stanford University Technical Report 474: Wavelets and Statistics Lecture Notes, 1995. – 27 p.

[Burbary, 2009] Burbary K., Cohen A. A Wiki of Social Media Monitoring Solutions // (on-line: http://wiki.kenburbary.com/)

[Burke, 2001] Burke M.M. Knowledge Operations: above and beyond Information Operations. 6th International Command and Control Research and Technology, June 19 – 21, 2001. – 16 p.

[Clauset, 2008] Clauset, A., Moore, C., Newman, M.E.J. Hierarchical structure and the prediction of missing links in networks // Nature, 2008. – 453, 98-101.

[Dean, 2004] Jeffrey Dean, Sanjay Ghemawat, «MapReduce: Simplified Data Processing on Large Clusters», декабрь 2004 года, http://labs.google.com/papers/mapreduce.html.

[DoD, 2003] Information operations roadmap – DoD US, 30 october 2003. – 78 p.

[Erdős, 1960] Erdős P., Rényi A. On the evolution of random graphs, Publ. Math. Inst. Hungar. Acad. Sci. 5, 1960. – P. 17-61.

[Ghemawat, 2003] Sanjay Ghemawat, Howard Gobioff, Shun-Tak Leung, «The Google File System», октябрь 2003 г., http://labs.google.com/papers/gfs.html.

[He, 2007] He B., Patel M., Zhang Z., Chang K. C.-C. Accessing the Deep Web: A Survey // Communications of the ACM (CACM), 50(5):94-101, 2007.

[Hill, 2000] Hill J.M.D., Surdu J.R., Ragsdale D.J., Schafer, J.H. Anticipatory planning in information operations // Systems, Man, and Cybernetics, 2000 IEEE International Conference, 2000. – 4. – P. 2350-2355.

[Kacperski, 2000] Kacperski K., Holyst J.A. Physica A. Phase transitions as a persistent feature of groups with leaders in models of opinion formation // Statistical Mechanics and its Applications, 2000. – 287, Issues 3-4. – P 631-643.

[Knight, 2003] Knight J.C., Strunk E.A., Sullivan K.J.Towards a Rigorous Definition of Information System Survivability // Proceedings of the DARPA Information Survivability Conference and Exposition (DISCEX'03), 2003.

[Lande, 2012] Lande D.V., Kalinovskiy Ya.A., Boyarinova Yu. E. The model of information retrieval based on the theory of hypercomplex numerical systems // Preprint Arxiv 1205.3031. (online: http://arxiv.org/abs/1205.3031).

[Lande, 2019] Dmytro Lande, Ellina Shnurko-Tabakova. OSINT as a part of cyber defense system // Theoretical and Applied Cybersecurity, 2019. - N. 1. - pp. 103-108.

[Lasswell, 1948] Lasswell H.D. The structure and function of communication in society // The Communication of Ideas. / Ed.: L. Bryson. – New York: Harper and Brothers, 1948.

[Latane, 1981] Latane B. The psychology of social impact // American Psychologist, 1981. – 33. – P. 343-356.

[Latane, 1997] Latane B., Nowak A. Causes of polarization and clustering in social groups // Progress in communication sciences, 1997. – 13. – P. 43-75.

[Lewenstein, 1993] Lewenstein M., Nowak A., Latane B. Statistical mechanics of social impact // Physical Review, 1993. – A, 45. – P. 763-776.

[Li, 2012] Li Y., Miller E.L., Long D.D.E. Understanding Data Survivability in Archival Storage Systems // Proceedings of the 5th Annual International Systems and Storage Conference (SYSTOR 2012), June 4–6, 2012, Haifa, Israel.

[Milgram, 1967] Milgram S. The small world problem, Psychology Today, 1967. – 2. – P. 60-67.

[Newman, 2003] Newman M.E.J. The structure and function of complex networks // SIAM Review, 2003. – 45. – P. 167-256.

[Nowak , 1990] Nowak A., Szamrej J., Latane B. From private attitude to public opinion: A dynamic theory of social impact // Psychological Review, 1990. – 97. – P. 367-376.

[Osgood, 1954] Osgood Ch. E. Psycholinguistics. A Survey of Theory and Research Problems // Supplement to the International Journal of American Linguistics. Vol. 20. No 4. Oct. 1954, mem. 10. Baltimore: Waverly Press, 1954.

[Price, 2001] Price G., Sherman C., Sullivan D. The Invisible Web: Uncovering Information Sources Search Engines Can't See. – Information Today, Inc., 2001. – 439 p.

[Roberts, 2002] Roberts P.W., Dowling G. R. Corporate reputation and sustained superior financial performance // Strategic Management Journal, 2002. – 23. – № 12. – P. 1077–1093.

[Shvachko, 2010] Konstantin Shvachko, Hairong Kuang, Sanjay Radia, Robert Chansler. The Hadoop Distributed File System. Proceedings of MSST2010, May 2010.

[Scaling, 2008] «Scaling Hadoop to 4000 nodes at Yahoo!», http://developer.yahoo.net/blogs/hadoop/2008/09/scaling_hadoop_to_4000_nodes_a.html.

[Sobkowicz, 2003] Sobkowicz P. Effect of leader's strategy on opinion formation in networked societies // Preprint Arxiv (online: http://arxiv.org/pdf/cond-mat/0311566)

[Schramm, 1974] Schramm W., D.F.Roberts (eds.) The Process and Effects of Mass Communication. Univ. of Illinoise Press, 1974.

[Watts, 1998] Watts D.J., Strogatz S.H. Collective dynamics of «small-world» networks. // Nature, 1998. – 393. – P. 440-442.

[Yahoo, 2008] «Yahoo! Launches World's Largest Hadoop Production Application», 19 февраля 2008 г., http:// developer.yahoo.net/blogs/hadoop/2008/02/yahoo-worlds-largest-production-hadoop.html.

Веб-сайты по тематике конкурентной разведки

1. Международное Общество профессионалов конкурентной разведки SCIP (www.scip.org)
2. Академия конкурентной разведки Fuld-Gilad-Herring, Кембридж (www.academyci.com)
3. Сообщество Практиков Конкурентной Разведки, СПКР (razvedka-open.ru)
4. Российское общество профессионалов конкурентной разведки, РОПКР (www.rscip.ru)
5. Институт конкурентной разведки, Германия (www.institute-for-competitive-intelligence.com/start.html)
6. Бизнес-школа Skema, Франция (http://www.skema-bs.fr/faculte-recherche/centre-intelligence-economique-et-influence)
7. Объединение профессионалов конкурентной разведки со штаб-квартирой в Канаде Competia (www. competia.com)
8. Харьковская областная общественная организация «Общество аналитиков и профессионалов конкурентной разведки» (www.scip.org.ua)
9. «Knowledge Camp & Competitive Intelligence Camp» – украинский BarCamp по конкурентной разведке и менеджменту знаний (barcamp2010.scip.org.ua)
10. Частная разведывательная компания «Р-Техно» (www.r-techno.com)
11. Агентство конкурентной разведки «Информант» (www.informnn.ru)
12. Технологии разведки для бизнеса «IT2B» (www.it2b.ru)
13. Альт-маркетинг: библиотека материалов по конкурентной разведке (alt-marketing.ru/articles/index-competitiveintelligence.shtml)
14. Информационная корпоративная служба (z-filez.info)
15. Конкурентная разведка, сайт Е.Л. Ющука (ci-razvedka.ru)
16. Конкурентная разведка в Интернете. Авторский курс А. Масаловича (http://www.tora-centre.ru/razvedka.htm)

Адреса упоминаемых веб-ресурсов

www.aignes.com – WebSite-Watcher – программа мониторинга веб-сайтов, форумов, локальных файлов.

www.anbr.ru – информационно-аналитическая система «Семантический архив».

www.archive.org – интернет-архив (Internet Archive).

archive-it.org – академическая поисковая система Biefield Билефельда BASE – одна из крупнейших в мире поисковых систем академических веб-ресурсов.

attackindex.com – система мониторинга информационных операций.

www.babkee.ru – Babkee – система мониторинга упоминаний в социальных медиа.

www.base.ukrpatent.org/searchINV – интерактивная БД «Изобретения (полезные модели) в Украине».

blog.capterra.com/top-8-free-and-open-source-business-intelligence-software/ – обзор средств бизнес-разведки.

books.google.com – Google Book Search – поиск книг.

brigh-tplanet.com – американская компания BrightPlanet, одна из первых опубликовавшая доклад о «глубинном веб».

www.ciradar.com/Competitive-Analysis.aspx – CIRadar – систем поиска информации для конкурентной разведки в «глубинном Интернете».

citeseerx.ist.psu.edu/index – CiteSeerX – электронная библиотека научной литературы и поисковая система.

www.data.gov – государственный сайт США, предоставляющий доступ к открытым государственным данным.

code.google.com – Google Code Search – поиск программного кода.

www.cy-pr.com – сервис анализа сайтов.

www.digitalpreservation.gov – американский национальный проект сохранения и распространения цифрового контента Digital Preservation.

www.dtsearch.com – dtSearch – поисковая программа, позволяющая обрабатывать статические и динамические данные во всех форматах MS Office.

www.eapo.org – Евразийская патентная база.

www.elastic.co – компания Elastic, разработчик трех связанных проектов – поисковой системі Elasticsearch, механизмом сбора данных и анализа журналов Logstash и платформой аналитики и визуализации Kibana .

www.europages.eu – Europages – Европейская бизнес-директория.

facebook.com – Facebook – крупнейшаяя социальная сеть.

www.fmsasg.com – программа визуализации связей и отношений Sentinel Vizualizer.

global.factiva.com – подразделение компании Dow Jones, занимающееся предоставлением доступа к деловой и аналитической информации через свои информационно-аналитические службы.

www.findlaw.com – FindLaw – каталог, содержащий список свободно доступных баз данных нормативно-правовых документов.

gephi.org –программа визуализации и анализа сетей и графов.

google.com – глобальная информационно-поисковая система Google.

google.com/alerts – Google-оповещения.

hootsuite.com – Hootsuite – многофункциональный сервис для работы с социальными медиа.

hrazvedka.ru – блог о разведывательных технологиях в бизнесе.

www.ibm.com/products/i2-analysts-notebook – система визуального проектирования структуры данных для хранения данных о различных персонах и организациях.

www.infongen.com – InfoNgen – агрегатор информации, настраиваемый на уникальные темы.

www.i-teco.ru – система управления досье X–Files.

infostream.ua – InfoStream – система контент-мониторинга веб-ресурсов.

www.integrum.ru – крупнейшая архивная база данных СМИ «Интегрум».

www.internetsec.com – служба Internet Securities, поставляющая бизнес-информацию от 16 тыс. источников.

inventionmachine.com – система Goldfire Research – система обработки контента глубинного веб.

www.iqbuzz.ru – IQBuzz – сервис для мониторинга социальных медиа с возможностью подключения по запросам пользователей новых источников.

www.kodeks.ru – информационно-поисковая система по российскому законодательству.

www.kribrum.ru – Крибрум – технология, позволяющая отслеживать и анализировать упоминания брендов, продуктов, услуг и т.п.

www.labyrinth.ru – российская база данных «Лабиринт», составленная на основе публикаций ведущих бизнес-изданий.

www.lexisnexis.com – крупнейшая в мире полнотекстовая онлайновая информационная система LexisNexis.

www.linkedin.com – LinkedIn – социальная сеть для поиска и установления деловых контактов.

www.livejournal.com – «Живой Журнал», ЖЖ, LiveJournal, LJ – платформа для ведения онлайн-дневников (блогов).

www.loc.gov – библиотека Конгресса США.

medium.com – платформа для социальной журналистики.

www.megaputer.ru – PolyAnalyst – семейство продуктов для глубинного анализа данных.

www.mlg.ru – «Медиалогия» – сервис, обеспечивающий онлайновый доступ к базе СМИ с возможностью производить самостоятельный мониторинг и экспресс-анализ СМИ.

mednar.com/mednar/desktop/en/search.html – бесплатная поисковая система в глубокой сети, ориентированная на медицину.

modusbi.ru – Modus BI – платформа для бизнес-аналитики, позволяющая собирать и визуализировать данные из различных источников, формировать отчетность и создавать прогнозы.

neticle.com/textanalysisapi/ – Neticle Text Analysis– технология извлечения информации из неструктурированных текстов.

newspapermap.com – Newspaper Map – сервис, объединяющий геолокацию и информационно-поисковую систему по медиаресурсам.

www.ntsb.gov – авиационная база данных NTSB Aviation Accident Database.

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

www.oracle.com/business-analytics/analytics-cloud.html – интегрированный комплекс аналитических инструментов компании Oracle.

patents.google.com – поисковая система от Google, которая индексирует патенты и патентные заявки.

www.peerindex.net – PeerIndex – сервис анализа социальных медиа, определяет размеры влиятельности компании.

photoinvestigator.co – сервис для извлечения метаданных и другой информации из фотографий.

www.politicalinformation.com — сервис поиска в 5000 отобранных веб-сайтах политической направленности.

www.postrank.com – PostRank – система глобального социального медиа-анализа.

www.rco.ru – RCO – система выявления фактографической информации из неструктурированных текстов.

www.rocketsoftware.com/products/rocket-folionxt – программа, позволяющая выявлять сущности, их взаимные связи и события, в неструктурированных текстах.

www.sap.com/sapbusinessobjects – Sap Businessobjects Text Analysis – программа, позволяющая извлекать информацию о десятках типах объектов и событий.

scholar.google.com – Google Scholar – поиск научных публикаций.

www.scip.org.ua – Общество аналитиков и профессионалов конкурентной разведки.

screen-scraper.com – программа, позволяющая автоматически извлекать всю информацию с веб-страниц, скачивать подавляющее большинство форматов файлов, автоматически вводить данные в различные формы.

www.semanticforce.net – SemanticForce – сервис мониторинга неструктурированных источников информации.

www.socialmention.com – Socialmention – платформа для поиска и анализа информации в социальных сетях.

www.softpedia.com – Website-Finder, программа, поиска веб-сайтов, которые плохо индексируются системой Google.

sphinxsearch.com – полнотекстовая поисковая система для больших данных.

telegram.org – кроссплатформенный мессенджер, позволяющий обмениваться сообщениями и медиафайлами многих форматов.

top100.rambler.ru – продукт, включающий в себя несколько сервисов: счётчик посещаемости веб-сайтов и система веб-аналитики, рейтинг русскоязычного сегмента Интернета и тематический каталог популярных ресурсов.

www.socialbakers.com –объединенная платформа маркетинга на основе анализа социальных сетей.

www.tora-centre.ru/avl3.htm – Avalanche – система интернет-мониторинга и конкурентной разведки.

www.trackur.com – Trackur – инструмент для мониторинга и анализа социальных медиа. Позволяет отслеживать, например, репутацию брендов.

trends.google.com – Trackur – инструмент корпорации Google, который показывает, как часто определенный термин ищут по отношению к общему объему поисковых запросив.

twitter.com – Твиттер – крупнейший сервис микроблогов.

ukrpatent.org –Украинский институт интеллектуальной собственности.

visual.ly – система поиска инфографики в веб-пространстве.

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

webground.su – Webground – интегратор русскоязычных новостей.

websvodka.ru – инструмент автоматического контроля изменений интернет-страниц.

weibo.com – Sina Weibo (кит. 新浪微博) – сервис микроблогов, запущенный компанией Sina Corp.

www.worldindustrialreporter.com/solusource – Global Supplier Directory by Solusource – веб-интерфейс для конкурентной разведки от компании Thomas.

worldwidescience.org – ***World Wide Sicence*** – глобальный научный портал, состоящий из научных баз данных и порталов.

www.yellowpages.kiev.ua – «Желтые страницы» Киева.

www.yahoo.com – глобальная информационно-поисковая система Yahoo!

www.yandex.ru – глобальная информационно-поисковая система Яндекс.

www.youscan.io – YouScan – система профессионального мониторинга социальных медиа.

youtube.com – крупнейший видеохостинг, предоставляющий пользователям услуги хранения, доставки и показа видео.

www.youtube.com/playlist?list=PL-9OTQQwXf2XuDGO_EIewUOpzUXLDDfcL – OSINT Academy - учебный курс – YouTube

zakon.rada.gov.ua – информационно-поисковая система по украинскому законодательству.

Наукове видання

НАЦІОНАЛЬНА АКАДЕМІЯ НАУК УКРАЇНИ
ІНСТИТУТ ПРОБЛЕМ РЕЄСТРАЦІЇ ІНФОРМАЦІЇ

ДОДОНОВ Олександр Георгійович
ЛАНДЕ Дмитро Володимирович
ПРИЩЕПА Віктор Володимирович
ПУТЯТІН Володимир Григорович

КОМП'ЮТЕРНА КОНКУРЕНТНА РОЗВІДКА

Монографія

(Російською мовою)

Київ, ТОВ «Інжиніринг», 2020

Підп. до друку 06.04.2021. Формат 60×84/16.
Наклад 300 прим. Замовлення № 104