Школа веб-разведки
Инструменты и источники

По некоторым оценкам, более 97 % критичной для бизнеса онлайновой информации невозможно найти с помощью традиционных информационно-поисковых систем. Однако существуют инструменты, «заточенные» под выполнение задач конкурентной разведки.

По меткому замечанию Андрея Масаловича (автора поисковой системы для конкурентной разведки Avalanche), из 23 видов поисковых задач, интересующих аналитика спецслужб, «Яндекс» удовлетворительно решает одну. В этом легко убедиться, набрав какой-нибудь интересующий вас запрос на любой из популярных сегодня поисковых интернет-систем — будь то Google, «Яндекс» или Rambler.

Поиск информации в Интернете без использования поисковых средств только путем просмотра отдельных сайтов носит выборочный и случайный характер (к тому же информация на отдельных сайтах может оказаться весьма субъективной, а порой и откровенно заказной) и крайне непродуктивен: вряд ли вы сможете обойти и просмотреть более десятка сайтов за день непрерывной работы, даже если теоретически знаете и помните их адреса.

Суммируя сказанное и перефразируя девиз одной из самых популярных российских поисковых систем «Яндекс»: «Найдется все!», можно сказать, что для конкурентной разведки «все» равнозначно «ничего», если не хуже. Поэтому правильным девизом для систем КР (CI) могли бы стать слова: «Найдется только то, что нужно, и ничего более».

Информация как хлеб насущный

По мнению бывшего директора ЦРУ Р. Хилленкерта, 80 % разведывательной информации получается из таких источников, как книги, журналы, научно-технические обзоры, фотографии, коммерческие аналитические отчеты, газеты, теле- и радиопередачи. Анализ только одного 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 «перелопачивать»: на протяжении года мониторилось 347 открытых источников, из них 295 — советских, а для создания одной страницы сводки ежедневно обрабатывался информационный массив объемом примерно 7 млн слов.

Итак, как мы выяснили ранее (см. ТЕЛЕКОМ 6/2007), открытые источники являются наиболее используемым каналом информации. А с ростом их количества, с одной стороны, возрастает объективность добываемой информации, но с другой — резко увеличиваются и трудозатраты на извлечение нужных данных. Следовательно, для их использования в конкурентной разведке нужны специализированные методики и системы.

И 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 в электронный вид, широкое использование и лавинообразное расширение Интернета, новые информационные технологии сделали аналитическую разведку в Сети одним из самых перспективных направлений разведывательной деятельности. А тот

факт, что таким образом поступают практически все спецслужбы мира, лишь подтверждает перспективность данного направления КР.

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

Как мы уже писали ранее, при ведении аналитической разведки в Интернете широкое распространение получило использование такого интересного направления науки, возникшего на стыке искусственного интеллекта, статистики и теории баз данных, как Knowledge Discovery (поиск знаний), использующего концепции Data Mining (добыча знаний в формализованных БД или потоках информации) и Text Mining (добыча знаний в полнотекстовых базах и информационных потоках). Уникальными особенностями Data Mining и Text Mining является то, что с их помощью можно вычленять из сырых данных ранее неизвестные, неочевидные, но полезные на практике и доступные для интерпретации знания, необходимые для принятия решений.

Одним из первых рассекреченных подобных комплексов стала французская система TAIGA (Traitement automatique d'information geopolitique d'actualite — автоматическая система обработки актуальной геополитической информации). Этот программный комплекс на протяжении 11 лет трудился в интересах французской разведки, после чего был заменен на более новый, рассекречен и разрешен к коммерческому использованию. Новый, более совершенный комплекс Noemic, взятый на вооружение французской разведкой, способен обрабатывать информацию со скоростью более 1 млрд знаков в секунду. Американский аналог этих программных комплексов Topic также рассекречен и передан в коммерческое использование.

Аналогичные системы разрабатывались и в бывшем СССР. Достаточно вспомнить такие из них, как «Барометр», «Эльбрус». Создание и использование подобных систем продолжается и сейчас — в России и других странах постсоветского пространства.

«Стоп! — скажет читатель. — Все вышеперечисленные системы либо используются государственными структурами, либо слишком дороги, чтобы их могли применять среднестатистические компании». На самом деле все не так плачевно. На современном рынке представлен целый ряд как западных коммерческих продуктов, так и продуктов производства стран СНГ, способных в том или ином объеме выполнять подобные задачи в интересах КР коммерческих структур.

Виды систем класса CI

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. Основная цель систем КР — информационное обеспечение перехода от традиционного интуитивного принятия решений на основе недостаточной информации к управлению, основанному на достоверных прогнозах и знаниях.

Основная цель систем КР — обеспечить переход от интуитивного принятия решений к управлению, основанному на достоверных прогнозах и знаниях

Безусловно, система конкурентной разведки, использующая Интернет как один из источников информации, должна настраиваться под специфику деятельности компании, включать в себя соответствующую классификацию, гибкие механизмы поиска, оперативной доставки данных, а также их каче-

ТЕЛЕКОМ-ИНФО

Недостатки традиционности

Поисковые системы отлично справляются с простыми однократными запросами. Но когда их нужно делать постоянно, предметная область сложна и слишком широка или, наоборот, предельно узка и отдалена во времени, а вам надо обобщить все информационные темы и поводы по данной тематике пусть даже за небольшой период, оценить их во временной динамике, найти пересекающиеся взаимосвязи с другими объектами, составить целостную картину об интересующем объекте, выделить нестандартное событие из общего массива фактов, то вы очень скоро убедитесь, что:

- ✓ поисковики либо перегружают вас миллионами бесполезных ссылок, либо, наоборот, ничего не находят;
- ✓ Интернет не хранит информацию долго, и то, что вы точно видели месяц назад на одном из любимых сайтов, сегодня можете там не обнаружить;
- ✓ поисковик не сохраняет просмотренные ссылки, и вам каждый раз приходится начинать титаническую работу с нуля после вынужденного перерыва;
- ✓ поисковик не отличает действительно важную информацию от шелухи;
- ✓ поисковик не умеет обобщать или сравнивать информацию по смыслу или другим критериям;
- ✓ поисковики принципиально не видят некоторых сайтов или отдельные виды информации на них (например, информацию, помещенную в БД), а некоторые сайты, наоборот, всегда отображают на первых страницах, хотя как раз они вас и не интересуют;
- ✓ поисковики могут выполнять поиск информации в Интернете только по вашему непосредственно введенному запросу и не могут повторять его автоматически в заданное время без вашего участия;
- ✓ вы привыкли пользоваться одним полюбившимся поисковиком, и хотя вы знаете, что результаты по одному и тому же запросу на разных поисковиках будут отличаться, у вас нет ни времени, ни желания их объединять.

ственной оценки. Одной из самых важных задач анализа информации является определение ее достоверности, то есть задача фильтрации шума и ложных данных. Без таких оценок всегда есть риск принять неверные решения. После анализа достоверности информации должны следовать оценки ее точности и важности. Главным критерием достоверности данных является их подтверждение другими источниками, заслуживающими доверия.

Информационные системы КР 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 отличается. Автоматически извлекаемые системой факты называют *А-фактами*, а факты, извлекаемые экспертами, — *Э-фактами*.

Существующие на рынке системы конкурентной разведки отличаются как по полноте и соответствию полному разведциклу, так и своему инструментарию и, как результат, цене. Кроме того, системы могут предназначаться для использования исключительно собственными силами внутреннего подразделения конкурентной разведки предприятия либо предполагать вынесение части задач на аутсорсинг специализированными структурами КР.

БД и аутсорсинговые отчеты

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 БД 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 структур, органов юстиции и судов, торгово-промышленных палат, органов приватизации и фондовых рынков, информационных, рейтинговых, аналитических и других агентств. Большую пользу приносят и отдельные доступные БД других контролирующих органов и организаций.

Традиционно КР опирается на такие источники, как опубликованные документы открытого доступа, которые содержат обзоры товарного рынка, информацию о новых технологиях, создании партнерств, слияниях и приобретениях, объявлениях о рабочих вакансиях, выставках, конференциях и т. п. Поэтому в последнее время все более популярны БД на основе архивов СМИ, в том числе и сетевых. В России, например, большой популярностью пользуются архивные базы данных СМИ «Интегрум» и «Медиалогия». В Украине эту нишу занимает, в частности, система контент-мониторинга интернет-прессы InfoStream, содержащая свыше 50 млн документов.

К разряду скрытого веба, например, относится и крупнейшая в мире полнотекстовая онлайновая информационная система Lexis-Nexis, которая содержит более 2 млрд документов с глубоким архивом до 30 лет по бизнес-информации и более 200 лет по юридической информации. Каждую неделю в архивы добавляется еще 14 млн документов. В отличие от неструктурированных массивов «поверхностного» веба Lexis-Nexis предлагает мощные инструменты поиска для получения достоверной и классифицированной информации.

Приведем еще один пример зарубежной БД из «теневого» веб-пространства. Компания ChoicePoint недавно предоставила сервис Auto TrackXP, вошедший в список двадцати круп-

ТЕЛЕКОМ-ИНФО

Объект под наблюдением

Основными объектами учета и мониторинга в системах конкурентной разведки, как правило, являются:

- ✓ источники информации (официальные сайты, интернет-издания, персональные сайты организаций или лиц, веб-представительства печатных СМИ, информагенств, теле- и радиоканалов, открытые базы данных 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 индивидуальные тематики.

ТЕЛЕКОМ-ИНФО

Западные системы бизнес-разведки

В последнее время все основные западные бренды, специализирующиеся на разработке хранилищ и баз данных, корпоративных системах управления, расширили свои линейки продуктов модулями Business Intelligence. О наличии таких модулей заявляют SAP, Oracle, SAS, IBM, Cognos и другие игроки.

Хорошо известна система Lotus Discovery Server от IBM — программный продукт, предназначенный для управления знаниями в корпоративных порталах. Система находит и идентифицирует связи, а также управляет интеллектуальным капиталом. Анализируя хранящуюся в организации информацию, Lotus Discovery Server может определять области экспертных знаний и подразумеваемые знания сотрудников, находя и организуя динамические связи между информацией, людьми и их деятельностью.

Среди самых развитых систем управления знаниями, применяемых для решения задач КР, нельзя не назвать систему Hummingbird Enterprise канадской компании Hummingbird. Среди множества компонентов системы можно выделить Hummingbird Portal — платформу, позволяющую интегрировать информацию из информационного хранилища и приложений в едином веб-интерфейсе. Эта платформа, как и ранее названный портал IBM Lotus, является полнофункциональным порталом знаний.

Еще одна флагманская платформа для конкурентной разведки — система Documentum (EMC), предназначенная для управления неструктурированной информацией, хранящейся в виде файлов различных форматов. Documentum основана на трехуровневой архитектуре, включающей хранилище содержания (репозитарий), службу управления содержанием (контент-сервер) и клиентские приложения для работы с контентом. Система позволяет хранить неструктурированный контент (веб-содержимое, XML-документы, мультимедиа-данные) и управлять им.

Для решения информационно-аналитических задач в настоящее время также широко используется система Cognos Business Intelligence, базирующаяся на идеологии OLAP. Одна из особенностей решения — возможность его интеграции с компонентами других ИС, в том числе необходимых для проведения бизнес-разведки систем финансово-экономического планирования и управления клиентской базой. В этом случае обеспечиваются широкие возможности сбора и консолидации данных из внутренних и внешних источников.

нейших скрытых сайтов мира (по рейтингу BrightPlanet). Auto TrackXP представляет собой базу данных объемом 30 ТБ, охватывающую почти все аспекты гражданской жизни США и содержащую информацию практически о каждом гражданине страны.

Testprofiles.com (часть ChoicePoint Online) содержит личные характеристики и сведения о компетентности граждан США. Например, чтобы определить, не завладел ли человек чужими документами, на основе системы организован платный сервис ProCheck, позволяющий сопоставить информацию из различных источников и государственных каталогов.

В Украине и других странах СНГ популярны такие базы данных, как российская БД «Лабиринт», составленная на основе публикаций ведущих бизнес-изданий (можно получить обширную информацию о конкретных персонах, организациях и компаниях), «Компасс», «Каре», «Желтые страницы», национальные представительства таких мировых брендов, как Dun & Bradstreet, Credireform, Europages и многие другие. Задача по поименному перечислению всех источников информации просто невыполнима, так как здесь должно действовать правило: чем большим количеством независимых источников подтверждается информация, тем более она достоверна.

Одним из самых эффективных источников информации могут служить отчеты и справки аутсорсинговых компаний, профессионально занимающихся КР и сбором сведений о коммерческих структурах и рынках.

В мире существует множество таких специальных компаний. Одной из крупнейших (ей принадлежит около 80 % западного рынка) является американская фирма, чья БД упоминалась выше, — Dun & Bradstreet. Справка по любой компании в этой службе будет стоить от $100. Серьезный анализ рынка или конкурента может обойтись от $10 тыс. Срок исполнения — от нескольких часов (информация присутствует в базе данных) до нескольких суток для справок и нескольких месяцев для серьезной аналитической работы.

В Европе не менее известны ирландская компания Credireform, немецкая Shufe, австрийская Intercredit, латвийская Coface IGK и др. Некоторые из этих фирм совмещают функции конкурентной разведки с другими видами деятельности, например обязанностями кредитных бюро. Другие специализируются лишь на КР.

Общей проблемой при обращении за справками в западные агентства, имеющие представительства в СНГ, является то, что, как правило, информация, предоставляемая в отношении западных нерезидентов, намного обширнее и качественнее, чем данные по отечественным фирмам. Поэтому в таких случаях целесообразнее обращаться к «родным» информационным компаниям — и дешевле, и качественнее.

В Украине существует целый ряд подобных компаний. Из известных авторам статьи можно назвать «Авеста-Украина», «Сидкон», Межбанковская служба безопасности «СКИФ» и другие. На российском рынке пользуются популярностью информационные отчеты компаний «Р-Техно», «Медиалогия», «Синс», «Интегрум», «Кронос-Информ» и др. Расценки российских фирм вполне сравнимы с западными.

Инструментальные средства КР

Около пяти лет назад по заказу Гарвардского университета российские разработчики из «Инфорус» создали систему Avalanche, которая в процессе поиска формирует модель предметной области в виде набора «умных папок», каждая из которых знает, что в нее должно попасть. Наполнением папок занимается специализированный робот, который запускается с компьютера «хозяина» и добывает только то, что было запрошено. Avalanche — одно из первых эффективных решений, использующих современные технологии глубинного анализа текстов.

На российском рынке, который близок нам по своей специфике, помимо упомянутой выше Avalanche представлено довольно много CI-инструментов. К заслуживающим внимания, с точки зрения авторов статьи, можно отнести: информационно-аналитические системы «Медиалогия», «Интегрум», «Тренд», «Семантический архив», «Аналитический курьер», «Астарта», «Галактика-Zoom», «Аналитик-2», Intellectum BIS, «Артефакт», информационно-программные комплексы «Арион», X-Files 2004, «Тренд», Cronos и т. д. На украинском рынке в этом сегменте представлены такие системы, как Web-Observer, «Сфера», Infostream, X-Scif и др.

Более подробно на анализе таких систем мы остановимся в следующей статье. Сейчас же хочется отметить, что не все из названных инструментов являются доступными и необходимыми ввиду их высокой стоимости или по ряду других причин. Вместе с тем отдельные задачи КР могут быть частично решены вполне доступными средствами. Использование новых подходов, а также открытых и относительно недорогих источников позволяет уже сегодня эффективно поддерживать принятие управленческих решений по многим направлениям бизнеса. ●

Дмитрий Ландэ,
Виктор Прищепа

ТЕЛЕКОМ-ИНФО

Требования к CI-системе

Одним из основных общих требований к CI-системе должно быть соответствие цикла обработки информации в ней классическому информационному разведциклу. То есть система должна самостоятельно или с участием оператора обеспечивать:

- ✓ выбор тематики и направлений развединтереса (целеуказание);
- ✓ выбор источников информации (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 КР необходимо анализировать данные из всех доступных источников, то крайне важным требованием системы 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. Последнее предполагает наличие у всех объектов учета атрибутов времени, места и источника данных, а также невозможность их безвозвратного удаления из системы с течением времени.