Инструментарий аналитика

В данном материале описаны системы мониторинга, извлечения фактов, построения связей на основе анализа неструктурированных текстов

В последнее время во всем мире и в Украине получают широкое распространение системы управления текстовой информацией. Это связано, с одной стороны, с возрастанием количества источников такой информации, развитием сети Интернет, в частности, ее новостного сегмента, блогосферы, оцифровкой литературы, а с другой — с увеличением спроса на аналитическую информацию, например, в таких сферах, как конкурентная разведка, маркетинг, PR. Интерес к рынку аналитических систем (средств поиска и аналитического обобщения массивов текстовой информации) объясняют также ростом законодательных инициатив, судебных рассмотрений, изменений в законодательстве, которые приводят к необходимости дополнительного сбора и осмысления информации. В частности, в области правовой информации такие системы получили название e-discovery (средства поиска информации юридической направленности). По мнению аналитиков Forrester Research, затраты на механизмы e-discovery возрастут с $1,4 млрд в 2006 году до $4,8 млрд в 2011 году.

В настоящее время существует ряд систем, которые выполняют отдельные функции, необходимые для построения комплексной системы мониторинга информационных ресурсов, извлечения фактов, построения связей на основе анализа неструктурированных текстов. Ниже приведен краткий обзор таких систем.

RCO

Сайт разработчика — *www.rco.ru*. Российская система, основная функциональность которой — выделение на основе анализа текстов на русском языке содержательных сущностей и связей между ними. Объективно является самой развитой в СНГ в этом направлении.

Система предоставляется в виде готового программного обеспечения, а также библиотек программ для разработчиков в Windows-среде.

Инструментарий разработчика: RCO Morphology Professional SDK. Поддерживает все возможности грамматического анализа любого слова русского языка: определение грамматических характеристик слова, приведение к нормальной форме, получение требуемых слово-

форм. Цена (1 процессор) — 278 000 руб., годовая поддержка — 61 160 руб. RCO Fact Extractor SDK — для разработки информационно-поисковых и аналитических систем, требующих лингвистического анализа текста на русском языке. Цены на этот продукт не разглашаются до момента заключения договора.

Готовый инструментарий аналитика: RCO KAOT — информационно-аналитическая система для работы в локальной сети на базе MS Windows и MS Internet Information Server, которая реализует функции «интеллектуального» анализа и поиска текстовой информации с поддержкой веб-интерфейса. RCO Fact Extractor — персональное приложение для Windows, которое предназначено для аналитической обработки текста на русском языке и выявления фактов различного типа, связанных с заданными объектами — персонами и организациями. Цена на одно профессиональное рабочее место составляет 104 000 руб., годовая поддержка — 22 880 руб.

Клиенты — Банк России, ОАО «Газпром», «Тюменская Нефтяная Компания», «МДМ-Банк», «Альфа-Банк», «Итар-Тасс», ФСБ.

Галактика-ZOOM

Система российской компании «Галактика». Сайт разработчика — *www.galaktika-zoom.ru*. «Галактика-ZOOM» — это система, основная функция которой — поиск информации в больших информационных массивах, а также выявление значимых слов и словосочетаний документа, отражающих его смысл, сравнение документов, обнаружение сходства, различия, аномалий изучаемых объектов. С помощью комплекса можно строить так называемый «информационный портрет» — совокупность ключевых слов, связанных с заданным запросом.

Система не обеспечивает мониторинга веб-ресурсов, но может охватывать поток, сканируемый из Интернета внешними модулями.

Система «Галактика-ZOOM» функционирует в архитектуре клиент/сервер на современных платформах Windows. Существуют русская и англоязычная версии системы. Базовая версия программного обеспечения системы «Галактика-ZOOM» в зависимости от функциональности стоит от $6000 до $20 000.

Клиенты системы «Галактика-ZOOM»: НТВ, РТР, ТВЭЛ, ЮКОС, а также Федеральная служба налоговой полиции, Федеральная служба безопасности РФ, Центральная избирательная комиссия РФ и ее региональные отделения.

Основная функциональность системы RCO — выделение на основе анализа текстов на русском языке содержательных сущностей и связей между ними. Объективно является самой развитой в СНГ в этом направлении

Семантический архив

Разработчик — компания «Аналитические бизнес решения». Сайт разработчика — *http://www.anbr.ru/*.

Информационно-аналитическая система «Семантический архив» представляет собой инструмент для создания интегрированного хранилища информации с возможностью хранения досье на объекты мониторинга, происходящие события, а также текстовые документы. Помимо этого система позволяет хранить информацию, импортированную из различных реляционных баз данных. Пользователи-аналитики имеют возможность искать информацию, выявлять взаимосвязи между объектами и событиями, генерировать аналитические отчеты. Система обладает такими возможностями, как загрузка документов с поисковых серверов, почты с помощью интернет-роботов (однако без форматирования документов); выделение упоминаний объектов, отношений и событий в текстовых документах; визуализация информации в виде семантической сети; создание текстовых отчетов и аннотирования статей. Включает как отдельный блок систему автоматизированной проверки физических и юридических лиц (ИАС «Проверка заемщиков»).

По мнению аналитиков Forrester Research, затраты на механизмы e-discovery возрастут с $1,4 млрд в 2006 году до $4,8 млрд в 2011 году

Система «Семантический архив» представляет собой клиент-серверное приложение, работающее с СУБД MS SQL Server 2000/2005 (ОС Windows). Существует интерфейс с CRM-системой Oracle Siebel.

Цена поставки системы зависит в основном от двух факторов: полноты комплектации и количества лицензий. Еще один фактор, влияющий на цену, — наличие компонента, который позволяет в пользовательском режиме менять структуру базы досье.

Многопользовательская базовая версия стоит $10—40 тыс., однопользовательская — в два раза меньше. Годовое техническое сопровождение составляет 15 % от стоимости купленного комплекта ПО.

Среди заказчиков — Счетная Палата РФ, Институт Общественного проектирования, Ситуационный центр

Санкт-Петербурга, Институт безопасности бизнеса, Альфа-банк, АКБ «Ак Барс», УГМК, ОАО «Атомредметзолото» и др. (всего 70 корпоративных заказчиков).

Медиалогия

Сайт разработчика — *www.mlg.ru.* Медиалогия — это система мониторинга и анализа прессы, ТВ и радио, электронных СМИ и блогов в режиме реального времени.

Возможности фильтров позволяют отбирать нужные пользователям сообщения из СМИ, сортировать сообщения мониторинга по важности. База СМИ — порядка 4000 источников, взвешенных по регионам и отраслям. Реализовано: поиск информации, выделение объектов (компаний, персон, брендов); оценка заметности публикации; оценка негатив/позитив; группировка по темам, выявление связей между объектами; сравнение объектов.

Клиенты системы — РОСНАНО, «Газпром», Правительство РФ, Минздравсоцразвития РФ, ВТБ, «Билайн», Минэнерго РФ, Минпромторг РФ, Российские железные дороги, Raiffeisen BANK и др.

Система функционирует в среде ОС Microsoft Windows (2000, XP, Vista).

Цена поставки ПО зависит от конкретного внедрения. Пример: бюджет совместного проекта учебных ситуационных центров в МГИМО составил около 53 млн руб. (август 2009 года).

Информационно-аналитическая система «Семантический архив» представляет собой инструмент для создания интегрированного хранилища информации с возможностью хранения досье на объекты мониторинга, происходящие события, а также текстовые документы

PolyAnalyst

Разработчик — компания Megaputer Intelligence Inc. (США). Сайт разработчика — *www.megaputer.ru*. Система PolyAnalyst позволяет решать проблемы прогнозирования, классификации, кластеризации, группирования объектов, анализа связей, многомерного анализа и интерактивного создания отчетов.

Система PolyAnalyst (и ее компонент — система TextAnalyst) обеспечивает лингвистический и семантический анализ текста, выявление сущностей, визуализацию связей, систематизацию документов, резюмирование и обработку запросов на естественном языке.

На базе PolyAnalyst компания Megaputer Intelligence Inc. Предлагает программу анализа деятельности конкурентов, которая загружает необходимые текстовые материалы, отбирает из них нужную информацию в виде отчетов.

Система PolyAnalyst реализована в виде клиент-серверного решения на платформе ОС Windows.

Информация по ценам: PolyAnalyst 5.0 полный пакет, лицензия на одно рабочее место: 147 623 руб.; рабочее место (PolyAnalyst Workplace): 8614 руб.; PA Text Analysis TA (английский язык): 11 584 руб.; PA Text Categorizer TC — каталогизатор текстов (английский язык): 6980,06 руб. Для полнофункционального решения необходимо еще примерно 30 модулей стоимостью в среднем 10 000 руб. каждый.

Megaputer Intelligence Inc. имеет более 500 компаний-клиентов по всему миру, среди которых: Национальный комитет безопасности перевозок США (NTSB), Electronic Data Services (EDS), International Air Transport Association (IATA), A State Medicaid Agency, A Police Department, Southwest Airlines и т. д.

«Галактика-ZOOM» — система, основная функция которой — поиск информации в больших информационных массивах, а также выявление значимых слов и словосочетаний документа, отражающих его смысл, сравнение документов, обнаружение сходства, различия, аномалий изучаемых объектов

RetriewalWare

В настоящее время программный продукт RetrievalWare принадлежит компании Fast (Microsoft Subsidiary) Search&Transfer. Сайт разработчика — *www.convera.com*.

Семейство продуктов RetrievalWare (RW) обеспечивает поиск и анализ информации по запросам на естественном языке. Информация может быть представлена как в неструктурированном виде, так и в формализованных базах данных, в локальной сети организации или в Интернете.

В RW реализована технология «нечеткого» поиска, ассоциативного поиска на основе семантической сети, возможность извлечения сущностей из текстов, что позволяет автоматически находить документы по терминам, связанным по смыслу с заданным запросом.

RW имеет возможность включать в хранилище данных информацию, представленную как в файловой системе, так и СУБД, почтовых системах и системах документооборота, индексировать удаленные хранилища данных. Это свойство RW позволяет создавать единое корпоративное информационное пространство. RW обеспечивает пользователю просмотр документов более чем в 250 форматах, среди которых как широко известные DOC, RTF, TXT, PDF, HTML, так и специфические форматы. В RW реализована функция кросс-языкового поиска. В настоящее время проводятся работы по созданию украинского семантического сервера.

Цены: RetrievalWare Server, первый процессор — 2 227 680 руб., дополнительный процессор — 2 227 680 руб., лингвистические процессоры, семантические и таксономические картриджи — первый процессор 222 768 руб.

Полная функциональность обеспечивается набором из еще около 100 модулей. Цены на техническое сопровождение не указаны.

Среди пользователей RW (всего их свыше 5000) — правительства России, США, Великобритании, Израиля, Польши, Чехии, Венгрии и Швеции; патентные ведомства — Швейцарии, Англии, США, Узбекистана и России; банки и компании — Worldbank, ЦБ России, Внешторгбанк России, Swiss Bank, Boeing Company, General Electric, Intel, Ford Motor Company, Visa International.

InfoStream

InfoStream — это система контент-мониторинга интернет-ресурсов, разработанная в украинской компании «Информационный центр «ЭЛВИСТИ». Сайт разработчика — *www.infostream.ua*.

На основе системы InfoStream построена одноименная технология, охватывающая по состоянию на декабрь 2009 года свыше 4000 информационных веб-сайтов, представленных на украинском, русском, белорусском и английском языках. Ежедневно с помощью технологии InfoStream сканируется свыше 60 тыс. новых документов.

Система InfoStream обеспечивает доступ к оперативной информации с единого интерфейса (по мере ее появления в веб-пространстве) в поисковом режиме с учетом возможного дублирования и семантической близости документов, языковых версий, размеров документов их цифровой насыщенности и т. д.; доступ к ретроспективному фонду, охватывающему около 80 млн. документов; поддержку аналитической работы в режиме реального времени: построение сюжетных цепочек, дайджестов, диаграмм встречаемости и таблиц взаимосвязей понятий, медиарейтингов.

Система InfoStream функционирует на серверной платформе под управлением ОС типа Unix (FreeBSD). Доступ пользователей к системе осуществляется через веб-интерфейс или по электронной почте в режиме подписки.

Стоимость минимальной базовой версии программного обеспечения системы InfoStream составляет 150 тыс. грн. Стоимость доступа к системе InfoStream в режимах онлайн и подписки по электронной почте составляет от 240 до 660 грн. в месяц.

Пользователями системы и сервиса на основе технологии InfoStream является свыше 500 государственных организаций и коммерческих структур, таких как СНБО Украины, Министерство экономики Украины, Антимонопольный комитет Украины, Национальный центр по вопросам евроатлантической интеграции Украины, Украинское национальное информационное агентство «Укринформ», Фонд «Возрождение», Ощадбанк, Брокбизнесбанк, Кредобанк, Райффайзен Банк Аваль, Морской транспортный банк и др.

Семейство продуктов RetrievalWare (RW) обеспечивает поиск и анализ информации по запросам на естественном языке. Информация может быть представлена как в неструктурированном виде, так и в формализованных базах данных, в локальной сети организации или в Интернете

Интегрум

Российское информационно-аналитическое агентство «Интегрум» *(www.integrum.ru, www.integrumworld.com)* предоставляет сервис по обеспечению пользователей необходимой информацией.

«Интегрум» не может рассматриваться как поставщик специализированного программного обеспечения, а исключительно как крупнейшая в России служба доступа к электронной информации. Информационные хранилища «Интегрум» содержат свыше 500 млн оцифрованных материалов из 10 000 источников, доступ к которым можно получить в режиме поиска (с помощью поисковой системы Artefact) и подписки.

Подписчики службы «Интегрум» — компании и госучреждения могут получать такие услуги, как поиск в СМИ — в прессе, на ТВ и радио; поиск компаний, в том числе финансовой и организационной информации; мониторинг СМИ (исследования медиасреды); бизнес-аналитика (с рекомендациями экспертов); консалтинг (маркетинговые исследования, бизнес-планирование, ТЭО).

Базовая стоимость подписки составляет 30 000 руб. в месяц. Подписчики «по умолчанию» получают возможность поиска в режиме онлайн.

Среди тысяч зарубежных клиентов службы «Интегрум» можно выделить такие: Британская библиотека, Публичная библиотека Квинса, Оксфордский университет, Шанхайский университет иностранных языков, Институт международных исследований (Монтерей), ООН, The Wall Street Journal, «Радио Свобода», ARD.

Медиалогия — система мониторинга и анализа прессы, ТВ и радио, электронных СМИ и блогов в режиме реального времени

Другие системы

Autonomy IDOL Server *(www.autonomy.com)* — одна из крупнейших в мире систем обработки неструктурированной, текстовой, аудиоинформации из разных источников с последующей ее обработкой, анализом и управлением. Autonomy охватывает решения для управления информационными потоками и связанными с ними рисками. Технология Autonomy позволяет автоматически разбирать смысловое значение неструктурированной информации путем использования математических алгоритмов сопоставления «паттернов» для определения основных концепций, содержащихся во фрагментах информации. Компания охватывает свыше 20 тыс. клиентов, в том числе British Telecom, France Telecom, General Motors, Reuters, BBC, British Airways, но пока плохо представлена на рынках России и Украины. Цены на программные продукты Autonomy соответствуют ценам ее основного конкурента — системы RetrievalWare.

Webscan *(www.webscan.ru)* — российская служба мониторинга интернет-ресурсов Webscan Technologies. В настоящее время сканируется около 1000 информа-

ционных веб-сайтов. Информация предоставляется пользователям по подписке на запросы. Базовая стоимость подписки составляет примерно $200 в месяц (при подписке на год). Стоимость подписки на один запрос колеблется от $10 до $120.

Fast Search & Transfer (FAST) *(http://www.microsoft.com/pathways/fast/)* — производитель систем масштаба предприятия для поиска, выборки и анализа информации в реальном времени. Продукты и технологии FAST обеспечивают интеллектуальную обработку больших массивов структурированной и неструктурированной информации.

Attensity suite *(www.attensity.com)* — технология извлечения информации из неструктурированных текстов. Она позволяет извлекать информацию, спрятанную в неструктурированном тексте, и переводить ее в структурированные данные, имеющие связи, которые могут быть проанализированы теми же методами, что и другие виды структурированных данных.

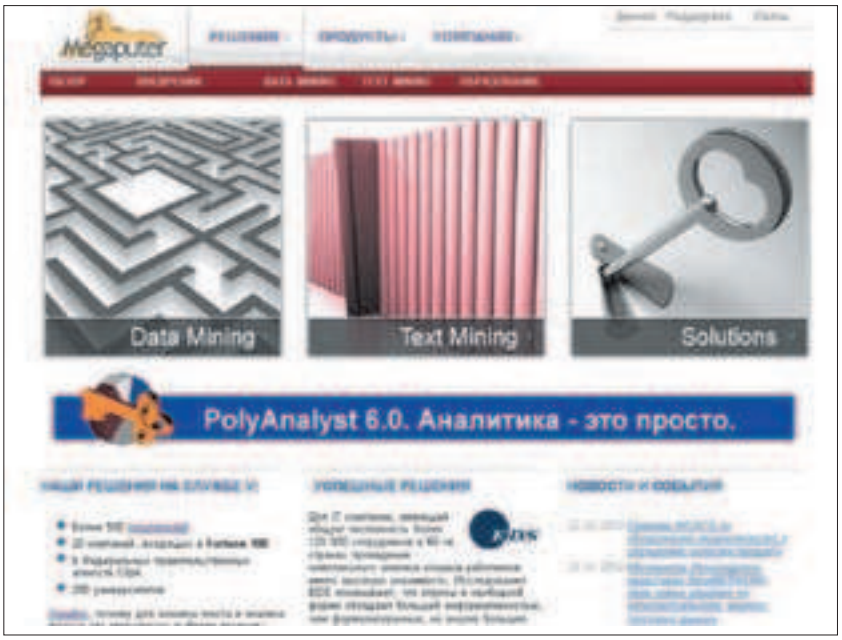
STATISTICA Text Miner — расширение программы STATISTICA Data Miner, предназначенное для перевода неструктурированных текстов в информацию, пригодную для принятия решений.

AeroText *(www.lockheedmartin.com/products/AeroText/products.html)* — программа, которая позволяет извлекать элементы информации, такие как сущности (entities), взаимоотношения (relationships) и события (events), в неструктурированных текстах, а также выявлять скрытые взаимосвязи и события в текстах.

На основе системы InfoStream построена одноименная технология, охватывающая свыше 4 тыс. информационных веб-сайтов, представленных на украинском, русском, белорусском и английском языках. Ежедневно с помощью технологии InfoStream сканируется свыше 60 тыс. новых документов

Businessobjects Text Analysis *(www.businessobjects.com/product/catalog/text_analysis/ features.asp)* — программа, которая, позволяет извлекать информацию по 35 типам объектов и событий, включая людей, географические места, компании, даты, денежные суммы, email-адреса, и выявлять связи между ними.

WordStat *(www.provalisresearch.com/wordstat/wordstat.html)* — программа, которая базируется преимущественно на статистическом анализе слов в слабоструктурированных текстовых документах и позволяет извлекать информацию из инцидент-отчетов, жалобных книг, обрабатывать результаты опросов, разрабатывать таксономии и др.



Система PolyAnalyst позволяет решать проблемы прогнозирования, классификации, кластеризации, группирования объектов, анализа связей, многомерного анализа и интерактивного создания отчетов

В настоящее время существует множество систем, которые могут применяться для задач аналитической обработки текстов, оказывать помощь в решении задач поиска, управления и обобщения массивов полнотекстовых документов. Самые мощные из таких систем имеют слишком большую стоимость с точки зрения отечественных пользователей. Большинство из названных выше систем совсем не адаптированы или плохо адаптированы к работе с документами на русском и украинском языках.

Вместе с тем, даже применяя отдельные доступные информационные системы, можно решать задание мониторинга, извлечения фактов, автоматического реферирования, построения связей на основе анализа плохо структурированных текстов. Приобретение больших систем не всегда необходимо, часто достаточно воспользоваться сервисными возможностями.

Развитие информационной инфраструктуры Украины, выход из кризисных ситуаций в экономике должны способствовать доступу специалистов-аналитиков к современным системам поиска и управления документами. ●

Дмитрий Ландэ,
зам. директора Информационного центра «Элвисти»