Современный вал информации в Интернете постоянно нарастает. Лидеры среди поисковых сервисов Рунета Google и Яндекс в ответ на это постоянно совершенствуют свои языки поисковых запросов. Однако усовершенствования касаются, по сути, повышения релевантности поиска словосочетаний и словоформ. Такие поисковые сервисы не могут искать связи между понятиями. Но, как известно, интерпретация понятия (факта) возможна лишь в связи с другими понятиями (см. об этом ИТМ №4 с. 10-12), и именно такая интерпретация важна в бизнес-разведке. Для решения такого рода задач разработаны технологии поиска связей между понятиями. О системах поиска на основе этих технологий речь идет в данной статье

Системы охвата информационных связей объектов мониторинга

Интернет: от поиска слов к поиску связей между ними

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

Поиск в неструктурированной информации

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

Неструктурированные тексты содержат в себе несравненно больше важной информации, чем структурированные записи баз данных, именно в силу того, что фор-

мализации подлежит сравнительно небольшой сегмент информации. В настоящее время появляется все больше качественных инструментальных средств извлечения фактов из неструктурированных текстов, таких как, например, RCO, Attensity suite, Businessobjects Text Analysis и т. д. В этом направлении ведутся активные исследования во всем мире. Следует отметить, что в рамках крупнейшей в СНГ конференции по компьютерной лингвистике «Диалог» (http://dialog-21.ru) в этом году проводился специальный круглый стол по тематике «Information Extraction», в работе которого приняли участие ведущие исследователи и разработчики из многих стран мира.

Рис. 1. Фрагмент страницы сайта Webground.
Правая колонка – «Темы дня» - информационный портрет

Факты и связи – что интерпретируем?

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 Сбербанка России с другими банками или частными лицами, то он не знает, какие банки или фамилии ему указать в запросе, а все документы, содержащие словосочетание «Сбербанк России», проанализировать физически невозможно. В таких случаях информационные связи, интенсивность которых выходит за рамки статистического фона, как правило, отражают реальность.

Интерпретируют обычно не сами понятия или факты, а взаимосвязи между ними. Важным оказывается не столько исследование самих понятий, сколько исследование их взаимосвязи. 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. Элементы такого подхода можно видеть, например, в «облаках» системы Quintura (http://quintura.ru), но там отображаются не понятия/сущности, а наиболее часто используемые слова. Подобные решения, реализованные в виде «информационных портретов», содержащих опорные слова, используются в таких системах, как «Галактика Zoom» (http://www.galaktika-zoom.ru), на веб-сайте интегратора новостей Webground (http://webground.su, см. рис. 1).

В системе контент-мониторинга InfoStream (http://infostream.ua) информационный портрет содержит не только опорные слова, связанные с запрашиваемой пользователем информацией, но и большое количество других сущностей, таких как фамилии персон, названия организаций, географические названия (топонимы), тематические рубрики и т. п. (см. рис. 2). Информационные портреты, используемые в названных системах реализуют граф в виде «звезды», центром которой выступает исходный запрос пользователей, а листьями – относящиеся к нему понятия. Однако на этом графе не отражается связь между отдельными понятиями – звезда не превращается в сеть. При этом объективно существует необходимость применения полнотекстовых информационно-поисковых систем, обеспечивающих поиск не по отдельным термам или понятиям, а по взаимосвязям между сущностями, присутствующими в документах.

Как искать связи

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 ребер теоретически может быть 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 v = 4 млн.

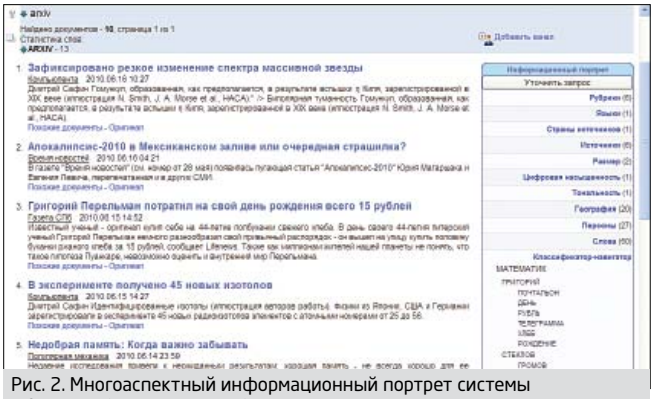

Рис. 2. Многоаспектный информационный портрет системы InfoStream (правая колонка)

Рис. 3. Место базы данных связей понятий в корпоративной информационной инфраструктуре

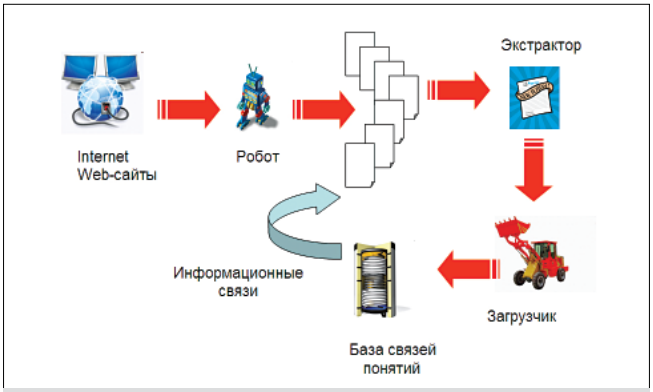

Рис. 4. Схема формирования базы данных связей

Кроме того, как показали эксперименты, распределение степени 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), должны 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 (например, http://peoples.ru, http://file.liga.net/person, http://openua.net), организаций(например,http://www.yellowpages.kiev.ua,http://ypag.ru, http://baza.kompass.ua), 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.

База данных связей

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. 3). 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.

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

На рис. 4 схематически представлены возможные технологические этапы формирования базы данных связей. С помощью программы-робота осуществляется сканирование выбранных веб-ресурсов, которые содержат информацию, относящуюся к объектам исследований. После этого осуществляется экстрагирование необходимых пользователям понятий, например, наименований брендов, компаний, электронных адресов и т. п. 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 Интернета, представленных на различных языках.

Технологии извлечения информации из неструктурированных текстов

Предложенный подход к поиску, естественно, влечет за собой некоторые особенности в реализации архитектуры базы данных связей понятий. 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 (http://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 сути, визуализацию связей, систематизацию документов, резюмирование и обработку запросов на естественном языке.
- Businessobjects Text Analysis (http://www.businessobjects.com/product/catalog/text_ analysis/features. asp) – программа, по-

зволяющая извлекать информацию о 35-типах объектов и событий, включая людей, географические названия (топонимы), компании, даты, денежные суммы, email-адреса и выявлять связи между ними;

- Attensity suite (http://www.attensity.com) – 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-Files украинской компании «Информационная корпоративная служба», которая позволяет пользователю в онлайн-режиме получать карты связей для выбраных им объектов и помогает интерпретировать результаты. Предусматривается, что пользователь вводит в качестве запроса системе объект. Запрос направляется к базе данных связей, откуда выбираются соответствующие ему фрагменты – карты связей (уровень детализации и временная ретроспектива должны указываться параметрически).

Граф информационных связей

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-Files 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. 5).

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. 5 приведен пример использования базы данных связей, случай, когда пользователя интересуют информационные связи Сбербанка России. Разумеется, для запроса «Сбербанк России» 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 являются банками-партнерами.

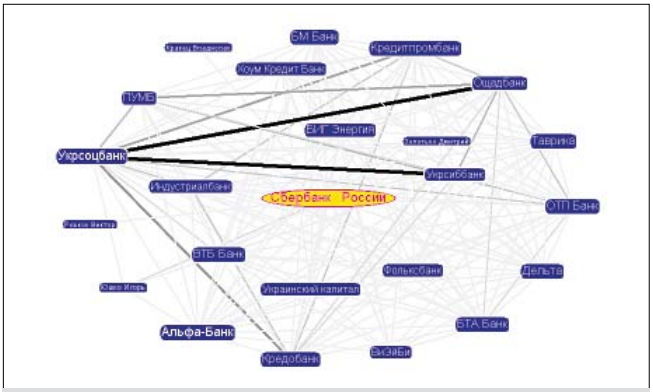

Рис. 5. Граф информационных связей понятия «Сбербанк России»

Перспективы «вертикального» поиска

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 используемых понятий и 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. Перспективы развития созданной системы – усложнение учитываемых связей, учет семантики контекста понятий в документах при их экстрагировании, отбор перечня действительно полезных баз данных текстовых документов, учет большего количества сущностей (понятий).

Представленный подход реализует связующее звено между полнотекстовыми и фактографическими базами данных. Очевидно, что реальный прорыв в области информационно-аналитической работы возможен лишь в результате агрегирования разных направлений. Базирующиеся на нескольких конкурирующих ранее точках зрения подходы на сегодня могут рассматриваться как пути создания современной мощной информационно-аналитической системы.

Дмитрий Ландэ,
Информационный центр ElVisti