ВАМ И НЕ СНИЛОСЬ, или Как приручить "девятый вал" информации

Д.Игнатьева, главный специалист ЗАО "МНИТИ" / ddi@mniti.ru
В.Филатов, к.т.н., с.н.с. ЗАО "МНИТИ" / filatov_vs@mniti.ru

УДК 004.9, DOI: 10.22184/2070-8963.2018.71.2.44.47

Сегодня в наш обиход стремительно входят новые понятия, такие как "Индустриальный интернет", "Интернет вещей" и др. Какую роль в этой индустрии играют системы интеллектуального мониторинга, каков расклад игроков в данном сегменте и что могут предложить российские разработчики?

Введение

По сути, сеть Интернет ("Всемирная паутина") в настоящее время является одним из ключевых драйверов развития мирового информационного общества, формирования национальных цифровых экономик. Россию пока трудно причислить к лидерам в области информатизации, однако и в нашей стране процесс формирования развитого информационного общества, проникновения интернета в различные сферы деятельности приобретает все большее ускорение. Принятая в прошлом году "Стратегия развития информационного общества в Российской Федерации", несомненно, даст новый импульс развитию различных цифровых технологий в нашей стране. Одним из основных направлений формирования цифровой экономики в России в этой "Стратегии..." названо направление обработки больших объемов данных, под которым понимается "совокупность подходов, инструментов и методов автоматической обработки структурированной и неструктурированной информации, поступающей из большого количества различных, в том числе разрозненных или слабосвязанных, источников информации в объемах, которые невозможно обработать вручную за разумное время".

Важность данного направления для российской экономики объяснять не надо. Автоматизация обработки больших объемов данных необходима во многих сферах деятельности: науке, промышленном производстве, транспорте и т.д. Особую важность это направление приобретает при решении различных задач с использованием сети Интернет. Казалось бы, что здесь может быть нового? Только ленивый не использует "Яндекс", Google и другие известные системы, осуществляя поиск необходимых сведений. Набрал в строке поиска название нужного товара, услуги, организации – и вся информационная паутина перед нами.

Правда, есть одно "но"... Этой информации очень много и, главное, она лавинообразно нарастает – удвоение объема информации сегодня происходит всего за несколько лет!

Проблемы поиска

Провести анализ информации, выбрать крупицу истины из вала различных новостей очень трудоемко, а зачастую невозможно. Анализ наличия спроса на тот или иной товар, оценка репутации интересующей компании, понимание того, как влияет географическое положение на мнение избирателей о будущем президенте своей страны, какие факторы оказывают на это влияние и как эти показатели меняются с течением времени – эти и многие другие задачи заставляют напрягать умы ведущих аналитиков многих компаний.

Каждому из нас не раз приходилось обращаться к "Яндексу", Google или другим поисковикам в надежде найти нужную информацию.

И каждый из нас помнит, как оперативно и без всяких проблем решался вопрос поиска требуемых данных, когда нам точно был известен адрес нужного сайта, или когда мы точно могли сформулировать интересующий нас вопрос. И каждый помнит, как трудно приходилось, когда интересующая нас проблема находилась на стыке различных понятий, или когда слово, вводимое нами в поисковик, имело распространенные аналоги. Например, при попытке выяснить стоимость моста, устанавливаемого взамен отсутствующих зубов в различных стоматологических клиниках, интернет упорно выдавал нам информацию о ходе строительства Крымского моста. Поиск данных для студенческого реферата о принципе работы доменных печей приводил к заваливанию нас данными о стоимости печей для дач, действующих специальных скидках и льготах при установке этих печей в течение ближайшей недели.

Не секрет, что в настоящее время сеть Интернет находит непосредственное применение не только в быту, но и для решения профессиональных (социологических, маркетинговых и др.) задач. Так, интернет наиболее оперативно и с наименьшими затратами может позволить оценить "раскрученность" компании по продаже средств гигиены, рейтинг различных претендентов накануне муниципальных выборов и т.д. При этом те же проблемы, с которыми сталкиваются обычные пользователи, характерны и для решения с использованием сети Интернет специальных задач. Более того, для получения требуемых оценок с высокой степенью объективности при проведении специальных исследований требуется еще обеспечить представительную выборку необходимых данных, т.е. выборку, характеризующуюся охватом различных сторон изучаемых объектов и наличием для каждого сегмента соответствующего объема данных. Понятно, что, например, маркетологам компании по продаже связного оборудования (смартфонов) или сотрудникам предвыборного штаба кандидата на должность главы управы решить такие задачи самим вряд ли удастся.

Справедливости ради нужно сказать, что разработчики поисковых систем прекрасно понимают эти проблемы и не сидят сложа руки. Например, представитель "Яндекса" в рамках форума CSTB-2018 рассказал о новых сервисах, предоставляемых его компанией, включая онлайн-трансляцию телевизионных каналов и интернет-кинотеатров. Но не это главное! Технология поиска информации становится все более "умной" и персонифицированной. Специальный "смарт-модуль" "Яндекса" теперь анализирует информационные запросы каждого пользователя и с учетом истории этих запросов проводит предварительный подбор информации с автоматическим отсеиваемой "лишней", по мнению поисковика, информации. Что по идее должно сделать процесс поиска и анализа информации более комфортным. Пробочным эффектом данной технологии станет адресная (таргетированная) рассылка рекламы каждому пользователю "Яндекса", формируемая с учетом истории его информационных запросов.

И это только первые шаги на пути автоматизации процедуры поиска информации, методы анализа которой уже разработаны и активно используются. Конечно же, в области анализа информации самый совершенный автомат в обозримой перспективе вряд ли сможет заменить человека, но появилась реальная возможность облегчить и сделать намного более эффективным труд аналитика: мониторить терабайты информации буквально за считанные секунды; объединять разные данные так, чтобы ими было легко оперировать. Сложившаяся ситуация сформировала социальный заказ на технологии, специализирующиеся на интеллектуальном мониторинге, т.е. на поиске в сети Интернет, каналах радио и телевещания, других средствах массовой информации больших объемов данных, их обработке и анализе. Иногда такие компании еще называют "датамайнинговыми" (Data Mining). И такие компании сразу же появились как на мировом, так и на российском рынке.

Большие данные: собрать, обработать, дать ответ

Наиболее продвинутой зарубежной "датамайнинговой" компанией является компания "Палантир", основанная в 2003 году. В 2011 году появился первый коммерческий проект: банк JPMorgan Chase заказал создание системы по борьбе с мошенничеством и контролю качества ипотечного портфеля по рекомендации полицейского управления Нью-Йорка, пользовавшегося продукцией "Палантир", – и к 2015 году доля коммерческих заказчиков в доходах компании составила 60%. Общая сумма инвестиций в проект превысила 3,32 млрд долл. США.

Основная концепция продуктов фирмы – обработка и визуализация больших массивов данных из разнородных источников, позволяющая

пользователям без технической подготовки находить взаимосвязи между объектами, обнаруживать совпадения между объектами и событиями вокруг них, выявлять аномальные объекты – Data Mining с упором на интерактивный визуальный анализ в духе концепции усиления интеллекта. В качестве источников программное обеспечение "Палантир" использует как традиционные базы данных и другие структурированные источники, так и тексты, а также аудио- и видеозаписи. При этом считается, что для непосредственного использования продуктов организациям-заказчикам не требуется персонал с инженерными или программистскими навыками, так как вся работа ведется в интуитивном графическом пользовательском интерфейсе, а запросы к источникам формулируются на естественном языке.

Другой крупной компанией, работающей на зарубежном рынке поиска и анализа больших объемов данных, является i2, которую в 2011 году для расширения спектра аналитических услуг и программных продуктов приобрела IBM. Решение задачи анализа связей и закономерностей в больших объемах данных реализуется в IBM i2 с помощью визуализации связей и статистических закономерностей между записями. Система IBM i2 состоит из различных компонентов, которые в зависимости от решаемых задач могут быть применены в различных вариантах архитектуры. Она позволяет собирать и анализировать данные из различных источников, представляя результаты анализа в виде схем взаимосвязей и временны́х диаграмм. На схемы также можно добавлять произвольные объекты, не содержащиеся в базе данных (например, текстовые комментарии или фотографии).

FuturICT – проект, разрабатываемый европейским консорциумом во главе с профессором социологии Швейцарского федерального технологического университета (Цюрих) Дирком Хельбингом. Проект официально запущен в 2011 году и рассчитан на 10 лет. Предполагаемая стоимость FuturICT – 1 млрд евро. Главной задачей проекта является построение информационной модели планеты. Это моделирование будет отслеживать все существенные события, от финансовых сделок, маршрутов миграции из одной страны в другую до данных о выбросе углекислого газа в атмосферу. Цель – сделать возможным моделирование будущего, подобное тому, как синоптики предсказывают погоду на основании выявленных скрытых связей. Смысл идеи не просто в диагностике. Реализованная модель даст возможность проигрывать различные сценарии будущего по типу "что если?" до принятия решений ответственными лицами.

Сервис Recorded Future позволяет аккумулировать информацию из более чем 150 тыс. различных СМИ, хранить архив до пяти лет с возможностью последующего анализа и извлечения знаний о вероятных последствиях произошедшего и будущих событиях. Инвестиции в эту компанию еще в 2009 году составили около 20 млн долл. США. Recorded Future занимается отслеживанием в реальном масштабе времени всего, что происходит в интернете; позволяет вычленять из контекста веб-страниц имена людей, места и действия, которые они упоминают; анализирует, когда и где те или иные события произошли. Затем искусственный интеллект пытается выяснить, как все это связано друг с другом. База данных, в которой записано около 100 млн событий, располагается на серверах Amazon.com, крупнейшего интернет-магазина в мире.

Российские компании интеллектуального мониторинга

Наряду с иностранными на российском рынке интеллектуальных поисковиков имеется несколько сильных отечественных компаний, предлагающих вполне конкурентоспособные решения.

Разработчиком одной из наиболее известных систем интеллектуального мониторинга "Катюша" является компания "М-13". В своей работе "Катюша" использует порядка 20 тыс. текстовых источников СМИ. Подаваемая пользователю информация проходит автоматизированную премодерацию (фильтрацию), вследствие чего аналитик получает отфильтрованное информационное поле, в котором отсутствуют лишние, по мнению эксперта-модератора, события.

Еще одним лидером рынка систем мониторинга и анализа СМИ является компания "Медиалогия", принадлежащая ООО "ИБС-Холдинг". Система мониторинга "Медиалогия" автоматически обрабатывает 500 тыс. сообщений СМИ и 50 млн сообщений соцмедиа в сутки.

Система мониторинга средств массовой информации "Скан" разработана в "Интерфаксе". Эта система нацелена на поиск в базе данных "Интерфакса" соответствующей информации, ее анализ и создание отчетов. Основные пользователи данной системы – те, кто "делают" новости и занимаются анализом медиасферы: главные редакторы, специалисты мониторинга,

журналисты, PR-специалисты, медиатехнологи, специалисты в области информационной безопасности и медиатехнологий.

Среди отечественных "датамайнинговых" компаний можно выделить и консорциум 3i Technologies. Он учрежден в 2014 году компаниями DSS Lab и InfoQubes, в 2016 к нему присоединилась компания PROMT. Консорциум объединяет российских разработчиков технологий, продуктов и сервисов для интеллектуальной обработки больших массивов разнородных данных и позиционирует себя на рынке облачной информационной поддержки бизнеса, делая основной упор на распознавание образов, анализ мультимедийных данных и аудиоархивов большого объема.

Среди российских компаний, работающих в области мониторинга информационных источников, обработки и анализа собираемых данных также можно отметить ЗАО "МНИТИ" (Московский научно-исследовательский телевизионный институт), разработавшее и предлагающее для коммерциализации одну из самых перспективных технологий интеллектуального мониторинга для поиска нужной информации, обработки, анализа и хранения собираемых данных.

Таким образом, системы интеллектуального мониторинга, разработанные российскими компаниями, также могут являться вполне надежной платформой для решения различных задач в области информационного обеспечения как телевизионных компаний, аналитических центров, так и других предприятий и организаций.

Заключение

Как видно из приведенного обзора, на российском рынке интеллектуального мониторинга и обработки больших объемов данных стали появляться отечественные и зарубежные компании, владеющие передовыми технологиями сбора и анализа информации и имеющие все необходимые составляющие для успешного продвижения. Несмотря на это, данный рынок в России до конца не сформирован и по уровню насыщенности (по объему и структуре) его можно отнести к категории дефицитного рынка. Одной из основных особенностей данного рынка является стопроцентная заинтересованность в нем всех предпринимателей, некоммерческих организаций и т.д., поскольку все прямо или косвенно участвуют в создании информации, а также являются ее потребителями. ■