информационные технологии

Бизнес в поиске

Проблему поиска данных и документов в корпоративных системах решают поисковые системы, разработанные специально для бизнес-задач. Пока их рынок в РФ невелик — 300–500 млн руб. в год. Но искать что-либо становится все сложнее: компаниям требуются все более интеллектуальные инструменты поиска, которые у российских разработчиков есть.

ЮРИЙ МАРТЬЯНОВ

Сотрудники российских компаний уверены, что массовые поисковые сервисы работают лучше корпоративных

— новые решения —

В это направление поиска по данным, накопленным внутри корпоративных систем, делают серьезные вложения крупнейшие компании — Facebook, Apple, Baidu и др. В России крупные поставщики и холдинги в текущем году тоже начали наращивать свои компетенции, стремясь быстрее ответить на запрос рынка. Rambler, например, приобрел компанию RCO, специализирующуюся на компьютерной лингвистике, информационном поиске и обработке неструктурированной информации. «Ростех» в лице Объединенной приборостроительной корпорации (ОПК) анонсировал готовность интегрировать разработки «Авикомп Сервисез», в частности лингвистический процессор Ontosminer для построения сложных систем текстового мониторинга и анализа данных, поддерживающий поиск не по ключевым словам, а по смыслу документа. Параллельно ОПК запускает масштабный проект в области искусственного интеллекта и семантического анализа, в котором участвует более 30 российских компаний, образовательных и научных организаций, в том числе ВШЭ и Бауманка.

Быстрого развития поисковых технологий требуют постоянный рост объемов данных и изменение структуры информационного пространства. «Поиск никогда не стоит на месте,— подтверждает Алла Забровская, директор по связям с общественностью Google в Центральной и Восточной Европе, России и СНГ.— В 2014 году мы внесли более 1 тыс. изменений, которые не всегда заметны пользователям, но тем не менее делают поисковую выдачу с каждым разом все лучше. Пока что система не дает ответов на сложные запросы, когда требуется объединить для ответа результаты трех или четырех поисков: покажи мне перелеты стоимостью менее 10 тыс. руб. туда, где в декабре жарко и можно заниматься дайвингом. Вот над решением таких проблем мы сейчас и работаем». При этом надо добиваться сокращения времени, которое потратит пользователь, обращаясь к системе с запросом.

Одно из важнейших направлений работы команды Google сегодня — это мобильные устройства и сервисы, так как заметно растет использование поиска именно с мобильных устройств. Важно также обеспечить возможность легкого переключения между разными устройствами без потери информации. Например, это удалось реализовать в Google. Фотосервис автоматически синхронизирует снимки и картинки со всех устройств, собирая их в альбом, поиском по которому можно пользоваться так же, как на десктопе.

«Яндекс», которому как поисковой системе в нынешнем году исполняется 18 лет, акцентирует работу с большими данными. Yandex Data Factory развивает услуги для компаний, нуждающихся в обработке больших массивов информации. «Машинное обучение, распознавание образов и речи, нейронные сети, обработка естественного языка — эти технологии „Яндекса“, используемые в YDF, выросли из первой экспертизы „Яндекса“ — поиска. Работа над ним не заканчивается никогда, а основная часть команды, работающая над поиском, сосредоточена в направлении поисковых сервисов»,— рассказывает Григорий Бакунов, директор по распространению технологий «Яндекса».

Нервы на пределе

Из относительно новых требований к поиску можно отметить сосредоточение на интересах конкретного пользователя, поиск по аудио- и видеоматериалам и социальный поиск. Это же относится и к корпоративному поиску: здесь растет интерес к классу решений, которые позволяют вести поиск в системах, подключаемых к внутренним базам, новостным сайтам, социальным сетям, форумам, тендерным площадкам и т. д.

Корпоративный поиск хотя и является неотъемлемой частью практически любой информационной системы, но до недавнего времени он оставался не самым очевидным явлением даже для бизнес-аудитории. Отчасти в силу своего недостаточного уровня развития, а также малой популяризации на фоне массовых веб-сервисов. По данным AIIM, около 70% респондентов считают, что найти корпоративную информацию гораздо сложнее, чем открытую в интернете. В российских реалиях две трети сотрудников компаний убеждены, что качество массовых поисковых сервисов выше, чем корпоративных, подтверждают результаты исследования 42Future.

«Поиск нужного документа иногда становится серьезной проблемой для корпоративного пользователя,— комментирует Олег Варламов, д.т.н., старший партнер и председатель научно-технического совета компании „Мивар“.— Часто проще запросить требуемую информацию у профильного отдела, чем выбрать из неструктурированной свалки разноформатных файлов. Из-за того что поиск нужной информации занимает слишком много времени, снижается производительность труда, сотрудники раздражаются, что ухудшает эмоциональный климат в коллективе».

Определенные трудности здесь создает и сама логика развития — и постоянного усложнения — корпоративной инфраструктуры. «Когда в компании много систем, поиск становится сложнее,— поясняет Алексей Сидорин, эксперт в области бизнес-аналитики и обработки больших данных компании КРОК.— Приходится искать в каждой и консолидировать результаты. При этом длительность поисковых проектов зависит от многих факторов: размера организации, ее баз данных, категории внедряемых поисковых систем, уровня кастомизации и пр. Реализация внутрикорпоративного поиска может занять в среднем от одного до двух месяцев, социальной сети — от двух месяцев до года, а системы бизнес-аналитики — от трех месяцев и больше».

Каждый ищет, как он может

Рынок корпоративных поисковых систем и многочисленных прикладных решений, основанных на поисковых движках (Search Based Applications), является частью концепции управления знаниями, тесно смыкаясь с технологиями текстовой аналитики (Text Mining), семантическими технологиями, технологиями оценки эмоциональной окраски, выявления мнений, автоматической классификации и многими другими, поясняют в компании «Преферентум» (ГК АйТи). Объем этого условно выделяемого сегмента в России пока невелик: эксперты из АйТи оценивают его в 300–500 млн руб. по предварительным итогам 2015 года. В мире речь идет уже о миллиардах долларов.

Глобально мощные поисковые платформы и инструменты для текстовой аналитики предлагает известная группа крупных вендоров. Например, IBM с продуктом IBM Content Analytics, Microsoft — с MS Fast. Активно развивает поисковые технологии SAP — в частности, в платформе SAP HANA. Здесь уже реализованы функции поиска с точным совпадением и с нечетким совпадением (fuzzy search) — когда при поиске может быть задан порог точности совпадения (к примеру, 90%) и в результате будут найдены все варианты, которые, по мнению системы, совпадают с искомым словом более чем на 90%. А также есть анализ текстовых документов (договоров, документации и т. п.) и выделение определенных сущностей (человек, организация, адрес и т. п.) и отношений между ними (например, данный человек работает в этой организации).

«Основная проблема, с которой сталкиваются разработчики,— необходимость обеспечения высокой скорости и точности работы поисковых запросов при работе с большими объемами и сложной структурой данных,— рассказывает Дмитрий Шепелявый, заместитель генерального директора SAP СНГ.— В учетных системах хранятся действительно большие объемы данных (десятки терабайт). А в связи с большим количеством и разнообразием автоматизируемых бизнес-процессов (финансы, логистика, сбыт, производство, различные индустриальные решения и т. п.) в системах хранятся документы разнообразной структуры и сложности».

Есть также примеры СПО-реализаций поисковых движков (на основе свободного программного обеспечения), например Apache Lucene/Solr, Sphinx, PostgreSQL Textsearch и т. д. Это полнотекстовые движки, поддерживающие множество языков, в том числе русский. Здесь как раз обнаруживается слабое звено: наилучшим образом с русским языком справляются все же отечественные разработчики.

Всего в России этим направлением занимается около 20 компаний, считают эксперты АйТи, подчеркивая, что многие отечественные разработки часто демонстрируют лучшую производительность и качество решения прикладных поисковых задач, чем технологии западных вендоров. «Некоторые продукты вообще не имеют аналогов в мире, например разработанная в АйТи система „Правовая экспертиза“, позволяющая выявлять правовые пробелы и коллизии, обнаруживать потенциальные коррупциогенные факторы в проектах нормативных правовых актов. Эта система несколько лет успешно работает в МВД и уже около года — в Государственной думе»,— рассказывает Дмитрий Романов, генеральный директор компании «Преферентум» (ГК АйТи).

При этом эксперты АйТи полагают, что в «чистом виде» корпоративная поисковая система в России пока не привлекает особенного внимания корпораций. В отечественных компаниях на нее трудно найти функционального заказчика, имеющего бюджет и готового его потратить.

Что умеет корпоративный поиск

В отличие от интернета, корпоративный поиск охватывает информационные системы с учетом прав доступа. Поиск происходит как на файловых серверах, так и на платформах (например, SharePoint или Exchange). Важно, чтобы корпоративный поиск учитывал особенности инфраструктуры, а также был интегрирован со всеми системами и мог индексировать разные форматы данных.

Современные тенденции в области поисковых средств расширяют само понятие поиска, считают в КРОК. «Помимо поиска всех объектов, связанных с запросом, актуальны также совместная работа и установка связей между объектами. В таких случаях используется система для совместной работы, например корпоративная социальная сеть как единая точка доступа ко всей информации: поиску сотрудников, документов, обсуждений, проектов, рабочих групп — всей доступной информации внутри компании. Пользователи могут не только оперативно находить нужные данные, но и работать с этим контентом из единого окна, а также оценивать контент и добавлять метаданные, что помогает поиску и категоризации»,— рассказывает Алексей Сидорин.

Работа большинства поисковых технологий основана на обработке больших данных, рассказывает Татьяна Даниэлян, заместитель директора по разработке технологий компании ABBYY. Как правило, поисковая выдача и ранжирование строятся на базе анализа статистики огромного количества взаимодействий пользователей и документов. Однако сотрудник, строя поисковый запрос, предполагает, что система сама поймет, в чем суть вопроса, найдет релевантные результаты и проранжирует их. При этом число обработанных поисковых запросов и последующих взаимодействий пользователей с результатами крайне мало по сравнению с аналогичной ситуацией интернет-поиска, а количество возможных неоднозначностей (омонимии, синонимы, пропущенные слова и т. д.) достаточно велико. Отсюда формируется потребность в поиске по смыслу, который основан на полном семантико-синтаксическом анализе документов и построении семантического поискового индекса. В результате поиск и ранжирование осуществляются по смыслу поискового запроса, а в выдаче человек получает релевантные документы, в которых могут быть и ключевые слова, и их синонимы — как обычные, так и смысловые.

«Например, „табурет“ и „стул“ не являются в чистом виде синонимами, но наша технология понимает, что они решают одну задачу, поэтому в этом ключе будут являться синонимами. Или если пользователь будет искать „положение об аттестации“, то сможет найти не только прямые совпадения, но и, например, «статьи о сертификации». Технология Compreno благодаря семантико-синтаксическому анализу настраивается на предметную область автоматически во время построения индекса»,— объясняет Татьяна Даниэлян.

Сыщики умнеют на глазах

Глобальная задача, стоящая сегодня перед всеми поисковыми системами, в том числе корпоративными,— это обеспечение возможности поиска по документу как единице поиска, считают в ABBYY. «Пользователи хотят формулировать свой запрос не просто в виде слова, фразы или предложения, они хотят на вход подать целый документ, а на выходе получить проранжированную выборку похожих документов. То есть в одной системе должны быть скомбинированы возможности поиска по словам, предложениям и документу целиком. Причем в идеале речь идет о поиске с запросом в виде документа, который может содержать и текст, и изображения и др. Предполагается, что будут и инструменты для ограничения получаемой выборки. Например, пользователь при поиске в такой системе по документу „приказ о назначении…“, может указать, что его не интересуют документы, которые относятся к финансовой части вопроса,— поясняет Татьяна Даниэлян.— Сейчас эта задача остро стоит для eDiscovery (процесс поиска информации в документах компаний в рамках юридических разбирательств, аудита и расследований), в научно-исследовательской области и в области безопасности. Сюда же можно отнести задачу поиска по сложным картинкам».

Будущее машинного обучения, которое используют большинство систем корпоративного поиска, связано с применением лингвистики и систем, основанных на семантике, считают в ABBYY. Это позволит при обработке входящих документов и поиске учитывать связи между словами в предложениях и на протяжении всего документа корректно распознавать омонимии и другие неоднозначности речи.

Переход к интеллектуальному поиску становится технологической тенденцией, уверены в «Миваре». «Первые попытки создать „осмысленный“ поиск уже предпринимаются. „Поиск 3.0“ отличается от привычного нам тем, что работает не с ключевыми словами, а с контекстами,— говорит Олег Варламов.— Системы должны научиться понимать, о чем идет речь и в каком смысле употребляется то или иное слово или выражение. Это необходимо, чтобы различать сходные по звучанию и написанию фразы, например „ключ“ как код и „ключ“ как инструмент».

Появление в поисковых сервисах контекстов потребует реализации принципов интерактивности — когда система, если не поняла смысла запроса, задает наводящие вопросы, а на следующем этапе, поняв контекст, начинает выдавать пошаговые рекомендации, считает эксперт «Мивар». На базе такой платформы можно будет реализовать принцип живой документации — когда ответ на запрос формируется в виде алгоритма из разных нормативных актов и инструкций. Или же интеллектуальные системы смогут проверять документацию на противоречивость и соответствие нормативной базе, реферировать корпоративные инструкции и документы по заданным параметрам, проверять почтовые сообщения на предмет разглашения конфиденциальной информации и т. д.

К семантическому поиску уже проявляют интерес крупные корпорации, например в США. Они ожидают, что его применение будет способствовать оптимизации и повышению эффективности их деятельности. Такие системы могут упростить и ускорить доступ сотрудников к информации, увеличить производительность труда, в том числе за счет роботизации процессов. В принципе в этом же должны быть заинтересованы и компании СМБ, которых сдерживает главным образом стоимость поисковых систем нового поколения. Ожидается, что начало их массового применения позволит достаточно быстро решить эту проблему — системы станут доступнее по мере их распространения.

Мария Попова