Зачем нам нужны технологии поиска и анализа неструктурированной информации?

Как оценить экономический эффект?*

*В предыдущих статьях мы рассмотрели подходы к анализу экономической эффективности СЭД и электронных архивов** и начали обсуждение вопросов экономической эффективности технологий поиска и анализа неструктурированной информации. Для упрощения под неструктурированной информацией далее в данной статье будем понимать прежде всего текстовую информацию – тексты документов, сообщения электронной почты, публикации в блогах и социальных сетях и т. п. Прочие виды неструктурированной информации – изображения, звук, видео и технологии для их анализа – оставим для отдельного рассмотрения в последующих публикациях.*

Д.А. Романов,
канд. физ.-мат. наук,
Национальный исследовательский университет «Высшая школа экономики»

Анализ, по определению, это разложение целого на составные части. Составными частями текста являются слова и предложения. Первое, что потребуется нам для разделения текста на составные части, – **технология, позволяющая извлекать текст из множества различных источников корпоративной неструктурированной информации** и понимать разнообразные и многочисленные файловые форматы, в которых сохраняют информацию используемые в организациях программные продукты.

Насколько объемным будет множество источников неструктурированной информации и насколько многочисленными окажутся файловые форматы в конкретной организации?

---

* Продолжение. Начало см.: *Романов Д.А.* Зачем нужны технологии поиска и анализа неструктурированной информации? Как оценить экономический эффект? // Современные технологии делопроизводства и документооборота. 2014. № 9. С. 21–30..

** См.: *Романов Д.А.* Как оценить экономическую эффективность системы электронного документооборота // Современные технологии делопроизводства и документооборота. 2014. № 1. С. 6–26; *Романов Д.А.* Зачем нужен электронный архив: как оценить экономический эффект? // Современные технологии делопроизводства и документооборота. 2014. № 5. С. 8–24.

Это зависит от размера организации и специфики ее деятельности, но в большинстве средних и крупных российских компаний можно смело рассчитывать на десятки (если не сотни) форматов файлов с текстовой информацией и не меньшее количество источников различных типов (от обычных файловых папок и реляционных баз данных до проприетарных форматов хранения текста в справочных правовых системах, CAD-системах*, на серверах MS Exchange и SharePoint, в базах данных Lotus Domino и т. п.). Большинство производителей решений для анализа неструктурированной текстовой информации реализуют средства для извлечения текста из различных корпоративных источников в виде специальных модулей – программных адаптеров, которые настраиваются на применяемые в организации информационные системы и извлекают из них текст для последующего морфологического анализа.

Среди базовых технологий, применяемых для поиска и анализа неструктурированной текстовой информации, можно назвать морфологический и синтаксический анализ текста, технологию, выделяющую из текста все упоминания информационных объектов, технологии обработки и выделения фактографической информации, определения степени похожести между текстами, классификации документов, анализа эмоциональной окраски и выделения мнений, анализа социального взаимодействия и другие.

Область деятельности, связанная с анализом неструктурированной информации, очень широка и разнопланова. И, безусловно, перечисленные технологии продолжают активно развиваться, в т. ч. исходя из практики конкретных программных продуктов, рассказ о которых читайте в статье.

Решения, использующие технологии анализа неструктурированной информации

Разнообразие технологий текстовой аналитики органично сочетается с еще большим разнообразием созданных на их основе программных продуктов и информационных систем различного масштаба и функционального применения. Множество программных продуктов используют данные технологии и с их помощью решают сложные задачи, еще несколько лет назад решавшиеся только человеческим мозгом либо вообще не решавшиеся, например, из-за огромного объема информации (на что у человека просто не было сил).

Мы все ежедневно сталкиваемся с технологиями поиска и анализа неструктурированной информации при использова-

* **CAD-системы** (*computer-aided design*) – компьютерная поддержка проектирования, предназначенная для решения конструкторских задач и оформления конструкторской документации (более привычно они именуются системами автоматизированного проектирования – САПР).

нии Интернета. Помимо собственно поиска, спектр применения таких решений простирается от контекстной рекламы, построенной на анализе ключевых слов из запросов пользователей и сыгравшей огромную роль в развитии поисковых серверов и Интернета в целом, до современных рекомендательных систем, позволяющих анализировать профили пользователя в социальных сетях, его покупки и поведение в интернет-магазинах и предлагать персонально ему подходящие товары и услуги.

В данной статье мы оставим за скобками многочисленные применения «персонального» характера и сконцентрируемся только на некоторых возможных способах корпоративного использования данных решений.

Корпоративная поисковая система

Это один из самых давних и хорошо известных вариантов применения рассматриваемых технологий. Проще говоря, если после приведения всех слов текста к их нормальной форме сохранить индексный массив (другими словами, записать на жестком диске информацию о том, какие слова, на каких позициях и в каких именно файлах находятся), а потом реализовать возможность по запросу предоставлять пользователю все файлы, в которых содержатся заданные ключевые слова, мы получим поисковую систему в ее простейшем, но вполне работоспособном варианте. Конечно же, реальные поисковые системы, рассчитанные на применение в ландшафте корпоративных систем, имеют значительно большие функциональные возможности. Прежде всего это относится к учету прав доступа пользователей к документам и файлам при проведении полнотекстового поиска.

На заметку!

Если в рамках какой-либо корпоративной информационной системы пользователь не имеет полномочий на просмотр документа, то и при полнотекстовом поиске этот документ должен оставаться недоступным для просмотра.

Кроме того, корпоративная поисковая система должна поддерживать технологии классификации информации, наглядно представлять результаты поиска, поддерживать поиск как по содержанию документов, так и по их атрибутам, интегрироваться с корпоративными порталами, системами электронного документооборота, электронными архивами и десятками других типов корпоративных информационных систем.

Система поиска экспертов

Другим интересным решением является система поиска экспертов. Не секрет, что в больших, территориально распределенных и быстро меняющихся организациях бывает непросто найти сотрудников с нужными компетенциями для участия в сложном проекте, выполнения НИОКР* по новой тематике,

* **Научно-исследовательские и опытно-конструкторские разработки** (англ. *Research and Development, R&D*) – совокупность работ, направленных на получение новых знаний и их практическое применение при создании нового изделия или технологии.

разработки инновационного продукта и т. п. Система поиска экспертов значительно упрощает задачу поиска нужных компетенций.

Она автоматически анализирует контент, создаваемый и получаемый сотрудниками организации (например, это может быть внутренняя электронная почта, личные страницы на корпоративном портале, научные публикации и т. п.), учитывает топологию и динамику информационных потоков в организации. В результате поиск эксперта становится еще проще, чем поиск нужной веб-страницы или документа, – в ответ на введенный запрос система предоставит список сотрудников (с их фотографиями, телефонами и адресами корпоративной электронной почты), наиболее релевантных заданной теме.

Вместо того чтобы изучать сотни документов, достаточно связаться с конкретным сотрудником и задать ему интересующий вопрос. Система поиска экспертов может иметь как внутрикорпоративную направленность (иначе говоря, искать нужные экспертные компетенции среди сотрудников организации), так и внешнюю, осуществляя поиск экспертов путем анализа открытой информации в Интернете: текстов монографий, статей в научных журналах, записей в блогах и форумах, выступлениях на конференциях и т. д.

Мониторинг СМИ и бизнес-разведка

Одним из давних и традиционных корпоративных применений технологий анализа неструктурированной информации является сбор, систематизация и анализ публикаций о компании, ее продуктах, проектах, акционерах, топ-менеджерах, конкурентах и т. п. Такой сбор сведений осуществляется для достижения нескольких целей.

Например, сотрудникам отделов маркетинга важно знать, какой информационный фон сопровождает проводимые маркетинговые мероприятия, как деятельность организации освещается в прессе, что потребители думают о ее продуктах и услугах. Информационный фон напрямую влияет на динамику курса акций компании, что не может не волновать финансистов. Инновационные продукты конкурентов являются предметом пристального интереса подразделения, отвечающего за исследования и разработки. Многим организациям необходимо комплексное управление репутационными рисками.

Для проведения такой работы и внедрения соответствующей масштабу задач информационной системы требуется использование практически всего спектра технологий анализа неструктурированной информации, рассмотренных в первой части статьи.

Подбор резюме для HR

Кадры, как известно, решают все. Управление персоналом – еще одна сфера применения решений, использующих технологии анализа неструктурированной информации. Трудно переоценить важность правильного подбора специалистов, особенно в тех видах деятельности, где знания и компетенции сотрудников являются основным капиталом организации.

Система управления персоналом, оснащенная технологиями анализа неструктурированной информации, автоматически сканирует сайты с описаниями вакансий и резюме соискателей, сопоставляет требования работодателей и компетенции специалистов, автоматически подбирает наилучшее соответствие.

Анализ корпоративной культуры, поиск информационных разрывов, выявление узких мест в бизнес-процессах

Это схожий с предыдущим вариант использования технологий анализа неструктурированной информации. Однако в данном случае ставится задача развития организации в целом, и анализ направлен уже не на отдельного человека, а на группы людей, информационные потоки между ними и способы их взаимодействия друг с другом.

Корпоративная культура – сложное многогранное понятие, одной из сторон которого как раз и является способ взаимодействия между людьми.

Для авторитарных организационных структур характерно иерархическое взаимодействие, передача управленческих решений сверху вниз и отчетов снизу вверх. Напротив, для демократичной культуры характерно наличие развитых горизонтальных связей, делегирование ответственности и полномочий, создание временных рабочих групп.

Анализ топологии, динамики и семантики информационных потоков позволяет выявлять в организации сотрудников, решающих схожие задачи, но не взаимодействующих друг с другом, автоматически обнаруживать реальную картину бизнес-процессов, находить «узкие места» в бизнес-процессах, выявлять неформальные сообщества и решать множество других важных задач организационного развития.

Правовая экспертиза

Система правовой экспертизы значительно упрощает процесс проверки проектов нормативных правовых актов (далее – НПА),

организационно-распорядительных документов, договоров и других документов.

С помощью такой системы сотрудник юридической службы организации за несколько секунд сможет установить следующее: не содержатся ли в проверяемом документе ссылки на нормативные правовые акты, которые утратили силу; нет ли в проверяемом документе фрагментов других документов, не возникает ли избыточное дублирование нормативной документации; соответствуют ли оформление и структура документа установленным в организации правилам; нет ли ошибок в договоре, соответствуют ли друг другу суммы, указанные цифрами и прописью, правильно ли рассчитан НДС и т. п. Система автоматически выделит в тексте документа упоминания о структурных подразделениях организации и проверит, соответствует ли этому списку лист согласования документа.

При обнаружении в проекте документа ссылок на внешние (российское законодательство) или внутренние документы (приказы, распоряжения, инструкции, регламенты, протоколы и т. п.) система сама сформирует гипертекстовые ссылки на них, и пользователю будет предоставлен быстрый доступ к тексту нужного документа (причем именно к той части, разделу или статье НПА, которая упомянута в тексте). Проанализировав заданный пользователем фрагмент текста, система выдаст перечень нормативных правовых документов, имеющих непосредственное отношение к заданной теме.

Важная возможность – автоматический анализ арбитражной практики и подбор дел с похожей правовой ситуацией.

Полезно знать

Интеграция с системой автоматизации бухгалтерского учета и (или) CRM-системой позволит обнаруживать и исправлять несовпадения в платежных реквизитах организаций контрагентов, а интеграция с реестром доверенностей поможет автоматически проверять, не истек ли срок действия полномочий у лица, подписывающего договор по доверенности.

Мониторинг торговых площадок

Для многих организаций существенной частью их коммерческой деятельности является участие в многочисленных конкурсах и тендерах, проводимых как государственными, так и коммерческими заказчиками. Для размещения объявлений о предстоящих конкурсах используются электронные торговые площадки. Однако количество таких площадок может исчисляться многими десятками (а количество конкурсов – сотнями и тысячами в месяц), и ручной мониторинг данной информации отнимает много времени.

Технологии анализа неструктурированной информации дают возможность автоматически отслеживать нужные торговые площадки в Интернете и своевременно информировать сотрудников организации о появлении потенциально интересной информации. Более того, современная система такого рода даже не нуждается в ручной настройке. Достаточно задать в качестве образца несколько десятков документов (например,

это могут быть технические задания, конкурсная документация на аналогичные конкурсы и т. п.), и это позволит системе самостоятельно определить профиль потенциальных интересов структурного подразделения организации.

Выявление плагиата

Повышение качества научных публикаций, диссертаций, результатов НИОКР, студенческих рефератов – важная и актуальная тема как на уровне государства в целом, так и на уровне отдельного предприятия – вуза, научного института, крупной компании. С применением технологий анализа неструктурированной текстовой информации (в частности, технологий, позволяющих измерять степень похожести между текстами) стало возможно реализовать эффективные системы для выявления некорректного заимствования текста и значительно снизить издержки и репутационные риски.

Современные системы выявления плагиата устойчивы к попыткам целенаправленной маскировки факта кражи чужого текста: перестановке слов, предложений, абзацев, замене слов на синонимы, добавлению «воды» и другим сильным искажениям чужого текста. Системы способны распознавать случаи корректного заимствования, такие как ссылки и цитаты. Другая интересная особенность – возможность исследовать степень семантической похожести между проверяемыми работами и выделять научные работы, статьи, публикации, составляющие отдельные тематические кластеры.

Важно, что окончательное решение о корректности или некорректности заимствования текста принимает эксперт – система только помогает ему в этом.

Автоматическая маршрутизация документов в СЭД

При автоматизации процессов документооборота возникает множество мест «принятия решений».

Например, при наложении резолюции на входящий документ принимается решение о том, кому расписать документ на исполнение. При отправке документа на согласование возникает необходимость выбора подходящего маршрута согласования. При проведении экспертизы ценности документов в архивах принимается решение, продолжать ли хранить документ в архиве или отправить его на уничтожение.

Система электронного документооборота, использующая анализ неструктурированной информации, может самостоятельно проанализировать содержание поступившего входящего документа и предложить перечень структурных подразделений организации, которые обычно занимаются исполнением документов с похожей тематикой.

При отправке на согласование внутреннего документа (например, проекта приказа) система сама выделит в тексте упомянутые наименования структурных подразделений организации и предложит включить их в лист согласования. Наконец, автоматический анализ содержания документов и выявление признаков, свидетельствующих о необходимости дальнейшего сохранения документа, поможет при проведении экспертизы ценности документов и определении сроков их хранения. Важно, что результатом работы системы являются не подмена человека и принятие за него решения, а рекомендации, позволяющие сберечь массу рабочего времени квалифицированных сотрудников.

Анализ жалоб, заявлений, обращений граждан

Другой интересной и полезной возможностью технологий анализа неструктурированной информации является автоматизация работы с письмами и обращениями граждан. Применение рассмотренных выше технологий для выделения из текстов обращений типовых информационных объектов (организации, персоны, адреса, тематика и т. п.) дает возможность за несколько секунд подготовить аналитический отчет, который вручную готовится много дней. Система сама проанализирует темы обращений, продемонстрирует распределение обращений на карте города или области, отметит изменения актуальности той или иной темы за заданный период времени.

Еще одной проблемой при обработке обращений граждан является обеспечение непротиворечивости в позиции организации. В больших и особенно в территориально распределенных организациях подготовкой ответов на обращения граждан занимается множество сотрудников и возникает риск того, что на одинаковые по смыслу обращения организация даст разные по смыслу официальные ответы. Для того чтобы устранить данный риск при подготовке ответа на поступившее новое обращение, система будет автоматически находить и предоставлять ответственному сотруднику все ранее поступившие обращения по схожим вопросам и подготовленные на них ответы.

Интеллектуальный корректор орфографии

Средства проверки орфографии давно и прочно заняли свое место в текстовых редакторах, клиентах электронной почты

и даже браузерах. Волнистая красная линия, выделяющая слова с ошибками, воспринимается пользователями как надежный и удобный инструмент, позволяющий быстро находить и корректировать ошибки и опечатки в текстах. Но помимо устранения простых опечаток и ошибок, необходимо обнаруживать и те, которые встроенные средства проверки орфографии не распознали.

Бывает, что все слова формально правильные, а вот их комбинация оказывается бессмысленной. Случается, что опечатка приводит к замене правильного слова на другое, «нормальное», вполне понятное и «честное» слово, но неуместное в контексте предложения. В этом случае типовые средства проверки орфографии не помогут и пропустят такую опечатку.

Например, если в предыдущем предложении ошибиться всего на одну букву и вместо словосочетания «честное слово» набрать на клавиатуре «честное олово», у текстового редактора не возникнет ни малейших подозрений. «Олово» так «олово», в словаре есть – значит, все в порядке, ошибка не обнаружена. И такая ситуация далеко не редкость.

На заметку!

Программа, оценив статистику совместного употребления терминов, проинформирует пользователя о возникновении потенциальной ошибки и предложит ему подходящие варианты замены.

Желающие могут лично убедиться в распространенности подобных ситуаций, поискав в Интернете «честное олово» или другие аналогичные ошибки. Интеллектуальный корректор орфографии, использующий технологии анализа неструктурированной информации, в таком случае заметит, что в тексте появилось словосочетание, которое не встречалось ранее, а встречалось очень похожее, отличающееся на одну-две буквы.

Управление подписками и новостными потоками, подготовка дайджестов

Не требует доказательств тот факт, что мы живем в век возрастающей информационной перегрузки. За неделю современный житель мегаполиса получает большее количество информации, чем его предок из XIX века получал за всю свою жизнь. Социальные сети, новостные порталы, традиционные средства массовой информации сражаются за внимание потребителя. Одна и та же информация может поступать к пользователю через разные информационные каналы, дублируя друг друга и отнимая драгоценное время. Полезная информация, действительно интересная и нужная пользователю, тонет под мегабайтами мусора, отвлекающего и рассеивающего внимание.

Технологии анализа неструктурированной информации позволяют вывести управление новостными потоками на новый уровень. Современные решения данного класса автоматически анализируют потребности и предпочтения пользователя, подсказывают ему другие интересные информационные ресурсы,

исключают дублирование новостных потоков. Можно сказать, что такое решение – это контекстная реклама «наоборот»: анализ поведения и информационных предпочтений пользователя осуществляется в интересах самого пользователя, а не внешнего рекламодателя.

Защита от информационных утечек

Контроль создаваемых и получаемых сообщений по электронной почте, анализ активности в социальных сетях и другие способы наблюдения за информационными потоками между сотрудниками организации представляют значительный интерес для специалистов по информационной безопасности.

Не секрет, что в крупных компаниях иногда сталкиваются с заметными потерями от деятельности так называемых инсайдеров – недобросовестных сотрудников, использующих в личных корыстных целях доступ к коммерческой информации.

Технологии анализа неструктурированной информации широко используются в DLP-системах – программных продуктах для предотвращения утечек конфиденциальных данных (аббревиатура DLP расшифровывается как *Data Loss Prevention* или *Data Leak Prevention*).

Итак, по результатам краткого рассмотрения программных продуктов можно сделать следующие выводы. Применяемые решения отличаются по нескольким признакам.

Во-первых, различие проявляется в характере индексируемых информационных ресурсов – внутренние, внешние, либо и те и другие.

Во-вторых, решения различаются по характеру использования и по кругу пользователей. Некоторые из решений представляют интерес и необходимы для поддержки деятельности всех сотрудников организации. Другие предназначены для автоматизации деятельности отдельных функциональных заказчиков и структурных подразделений. Третьи вообще не являются самостоятельными независимыми приложениями, а используются как сервисы другими корпоративными информационными системами. Такое разнообразие программных решений и способов их использования обусловливает и разные подходы к оценке экономической эффективности. Оценить экономический эффект от применения рассмотренных в статье программных решений вы сможете уже в следующем номере журнала.