ПРОБЛЕМЫ КОНТЕНТА РЕСУРСОВ ИНТЕРНЕТА

А. Г. Шаров, канд. техн. наук
ООО "МегаВерсия", Москва

Многообразие информации в Интернете, имеющееся в настоящее время, на первых этапах его изучения вызывает у пользователей впечатление, что в Сети можно найти практически любую необходимую информацию, ответ на любой вопрос, сведения из любого раздела жизни.

Этот поверхностный взгляд имеет под собой основания: в Сети, как показывает анализ, действительно имеются публикации, информационные материалы по огромному числу вопросов, но в то же время значительная часть этой информации, как правило, либо быстро устаревает (имеет низкую актуальность), либо носит непрофессиональный характер.

Не нужно забывать и такой фактор как достоверность информации. Свободное обращение информации в Сети привело к появлению множества публикаций, носящих явно дезинформирующий характер, содержащих сведения, противоречащие морали, этике и здравому смыслу. Это явление вывело на свет множество лженаучных теорий, негуманных либо противоправных публикаций и т. п.

Опубликованные сведения, их информационная ценность для различных категорий пользователей неоднозначна, что невольно заставляет разработчиков поисковых систем структурировать эту информацию.

Существующие поисковые системы пока достаточно успешно справляются со столь огромными объемами, перекладывая заботу об информационной ценности на плечи самих пользователей, давая им на откуп решение вопроса о пертинентности информации.

Перечень из многих тысяч сайтов и документов, найденных поисковой системой и предлагаемых в качестве результата поиска той или иной информации, заставляет задуматься о разработке таких механизмов управления контентом, его доставкой потребителю, которые действительно помогут снизить затраты человека на получение информации, отсечь заведомо ненужную либо неадекватную информацию.

Появление в последнее время "интеллектуальных" систем поиска, проводящих семантический (смысловой) анализ запрашиваемой пользователем информации, позволило снять часть нагрузки с пользователей, обеспечив им предварительную кластеризацию и представление результатов поиска в виде "деревьев" либо "облаков" терминов и понятий, близких по смыслу к запросу.

В то же время получение и использование информации из Сети, доставка сведений профильному потребителю остается наиважнейшей задачей разработчиков сервисов управления доставкой контента Интернет.

В качестве примера можно озвучить тезис о том, что развиваемая целевая программа "Образование" нуждается в таких средствах доставки, которые могут способствовать применению сервисов Интернета в учебных, образовательных целях. Не секрет, что "раздача компьютеров" по школам, создание компьютерных классов, локальных сетей и "подключение к Интернету" (чем, собственно, и закончилась данная программа) не решают вопроса получения учениками информации образовательного назначения — учебных, дидактических материалов, справочников, пособий, энциклопедий и т. п. Школы, получив свободный доступ к ресурсам Сети, не могут корректно управлять информационным потоком в силу того, что не имеют специалистов по управлению информацией. В результате вся нагрузка по контролю над учениками, работающими в Сети, ложится на преподавателей.

Одним из решений проблемы доставки необходимой информации профильного (образовательного) назначения целевым потребителям может стать фильтрация контента, получаемого из Интернета. В настоящее время существует система, которая уже на этапе поиска запрашиваемой информации вводит ограничения на доступ к ресурсам, содержащим ненужную либо неадекватную целевым потребностям информацию.

Введение такой, пока предварительной и весьма жесткой системы фильтрации контента позволяет решать сразу несколько насущных задач:

- первой прагматической задачей является сокращение потребления трафика и, соответственно, снижение нагрузки на телекоммуникационную сеть, а также уменьшение за счет этого расходов на оплату услуг провайдеров;
- одновременно решается вторая, более важная на мой взгляд, социальная задача. Интернет используется как образовательная информационная среда, а не как место для развлечений и получения несанкционированной информации, будь то электронная почта, получение мультимедийной информации и др.;
- третьей задачей является отсечение (фильтрация) нежелательных информационных артефактов — сайты националистической, агрессивной и террористической направленности, просмотр "веселых картинок", рекламные материалы, сведения о нелегальных организациях и сектах, доступ к противоправной информации и т. п. Все это имеет выраженную социальную (идеологическую) направленность на воспитание молодежи, обеспечение здорового морального и психического состояния детей и подростков, всей учащейся молодежи.

Естественно, степень (качество) фильтрации, уровень отсечения зависят от потребностей групп пользователей, для которых производится фильтрация. Для каждой из групп формируются правила доступа — так называемый "тематический профиль", в котором указаны разрешенные и запрещенные к использованию ресурсы Сети, календарь применения правил, другие необходимые для управления доставкой контента данные.

Если говорить о "школьном Интернете", то такими критериями могут быть списки ресурсов, относящиеся к конкретному разделу (в терминах информационного поиска — рубрике), изучаемой на данном занятии дисциплины.

Ограничения на доступ к информации должны носить разумный характер, поэтому системой автоматически производится классификация сайтов по тематической направленности, проводится экспертная оценка качества классификации, исключение информационного "шума", и на данной основе формируются рубрики. При этом в тематический профиль записываются сведения не только о данной предметной рубрике, но и о смежных рубриках, расположенных выше либо ниже по структуре классификации, наподобие "деревьев" и "облаков", использующихся в поисковых системах.

Таким образом, ученики на занятии имеют возможность изучения в интерактивном режиме Сети как программного учебного материала, ссылки на который находятся в тематическом профиле в качестве основных информационных ресурсов, так и получения дополнительных сведений, по смежным учебным дисциплинам, из справочников, энциклопедий и т. д.

При проектировании и реализации системы использованы наиболее современные средства извлечения знаний из информационного многообразия Интернета — сетевые агенты, технологии Omnibase, предусматривающие формирование запросов к внешним поисковым системам, алгоритмы поиска и обхода сайтов в Интернете для получения профильного контента.

Аналогичные проблемы есть и при профессиональном, профильном использовании контента Интернета. Бизнес стремится использовать информационное пространство Сети с максимальной выгодой, отсекая все ненужные ему артефакты стороннего характера.

В настоящее время активно развивается направление фактографического анализа, добычи данных и поиска конкретных фактов, которые могут повлиять на ведение бизнеса. Существует определенное множество систем, обеспечивающих такое использование информации Сети. Однако большая их часть представляет собой универсальные программные оболочки, предназначенные для создания на их основе пользовательских систем добычи и анализа профильной фактографической информации. Как показывает практика, внедрение такой профильной фактографической системы требует значительных временных, технических и ресурсных затрат.

Снижение затрат на разработку, адаптацию к профессиональным потребностям пользователя и внедрение систем фактографического анализа информации возможно за счет использования существующих в настоящее время массивов фактографической информации, сертифицированных программных средств (Oracle, RCO и пр.) и стандартизированных оболочек (по типу IE, Netscape Navigator, Opera и др.).

Снижение стоимости владения, сопровождения и эксплуатации системы фактографического анализа достигается за счет:

использования технологии контентной фильтрации, уже на первом этапе поиска в Сети ограничивающей объем информационных ресурсов;

применения адаптивных алгоритмов извлечения данных из информационных источников, позволяющих автономно сканировать информационные ресурсы для обнаружения появления в Сети информации, соответствующей сформированной потребности пользователя. Они основаны на использовании технологии мобильных агентов, настраиваемых (адаптирующихся) в соответствии с информационными потребностями пользователя;

использования "интеллектуальных" механизмов для поиска необходимой информации, позволяющих пользователю формировать запрос (в данном контексте точнее вопрос) на естественном языке, в виде произвольной фразы либо фрагмента текста. Дальнейшее развитие системы предусматривает использование механизмов, аналогичных применяющимся в развиваемой сейчас идеологии "семантического Веба", основанного на технологии WordNet и интеллектуальных методах анализа текстовой информации.

Нельзя также исключать ставшую уже традиционным приемом кластеризацию ответов информационно-поисковой системы с возможностью последующего уточнения вопроса либо области поиска необходимой информации.

Основанная на данных принципах система поиска, семантического анализа и рубрикации интернет-ресурсов (СПАР) в настоящее время уже реализуется.

Дополнительным преимуществом такого подхода является возможность построения на изложенных принципах информационно-аналитической системы для управления рисками различной природы (финансовыми, техническими, технологическими, кадровыми, образовательными). Введение в структуру системы модуля венчурного анализа (финансовых, технических, технологических и иных рисков) имеет шанс стать инструментом для бизнес-аналитиков в различных предметных областях для формирования прогнозов развития отдельных ситуаций, процессов и явлений. Бизнес требует наличия систем, способных не только предоставлять текстовую информацию (факты), но и производить ее автоматизированный анализ.

Особо актуальной задачей здесь является быстрое реагирование информационной системы на оперативные сведения, которые могут поступать из Сети, например, новостные ленты RSS, оперативные сведения по курсам и котировкам, хроника горячих событий и т. п. Как уже упоминалось, технические возможности разработки таких систем существуют.

Наличие методик KDD (Knowledge Discovery in Databases — обнаружение знаний в базах данных) и KDMine (Data Mining and Knowledge Discovery — добыча данных и обнаружение знаний), активно продвигаемых в настоящее время на западном рынке современных информационных технологий, позволяют надеяться на успешный исход работ в данном направлении.

Использование и практическая реализация этих методик позволят создать информационно-аналитическую систему, оперирующую фактографической информацией в форме числовых данных, отдельных характеристик и свойств объектов, для последующего ретроспективного анализа и построения прогнозных показателей для выбранного объекта или процесса.

Таким образом, содержание информационного поля Интернета и связанные с ним проблемы извлечения профессиональной информации для целевых категорий пользователей, обработки и доставки контента необходимой потребителю информации требуют разработки принципиально новых подходов к созданию информационно-аналитических систем. Время простого извлечения фактов из больших массивов информации прошло. Наступила пора активного использования существующих механизмов для создания структур фактографической информации об исследуемых объектах – баз данных и баз знаний, которые в свою очередь будут являться фактографической основой для работы аналитических систем в различных областях деятельности — бизнесе, политике, образовании и т. д.