Введение

Количество информации, обрушивающейся на человека в современном мире, обусловливает актуальность задачи отделения действительно важных сведений от информационного шума. Человек, группа людей, информационная служба, профессиональные эксперты-аналитики уже не могут пропускать через себя потоки информации, которые изливаются на них сегодня электронными медиа. Зачастую даже опытные эксперты не могут выделить главного, не находят сведений, необходимых для принятия решений, в результате чего действия как отдельных людей, так и коллективов или даже государств становятся неадекватными реальной обстановке.

Таким образом, самая главная проблема современных коммуникаций — это извлечение действительно ценных сведений из информационных потоков; другими словами, получение знаний из информации.

Обилие информации уже давно воспринимается как нечто само собой разумеющееся. Количественные оценки ее суммарного объема как таковые вряд ли могут стать поводом для особых размышлений. Но если подобные показатели подвергнуть структурному анализу, то полученные результаты могут оказаться весьма неожиданными.

Возьмем, к примеру, исследование изменения объема информации в мире за год [54]. С 2000 года оно проводится в Калифорнийском университете в Беркли под руководством профессоров Питера Лаймана и Хола Вэриена. Ученые пришли к выводу, что на протяжении трех лет, предшествующих 2002 году, количество информации, произведенной человечеством, удвоилось. А в самом 2002 году в мире было произведено пять экзабайт (миллионов терабайт) информации. Для сравнения приведем данные об объеме фонда библиотеки Конгресса США, где хранится 19 млн книг и 56 млн рукописей: он составляет около десяти терабайт информации. В упомянутом исследовании информация структурировалась по типам носителей. Оказалось, что лидерство прочно удерживают магнитные носители, доля которых превышает 90%. Из них большую часть составляют жесткие диски. На кино, фото, печатные издания и другие бумажные документы вместе с оптическими цифровыми носителями приходится лишь 7% информации.

Очевидно, что лишь человеческого опыта в данной информационной ситуации становится уже недостаточно. Сама среда поступления информации определяет и возможные реальные подходы к ее обработке. Только мощные возможности информационной техники — компьютеров, сетей — в совокупности со специальным программным обеспечением могут оказаться той панацеей, которая спасет нас от информационного хаоса. В свое время казались очень перспективными системы искусственного интеллекта, экспертные системы со своими парадигмами фреймов и правил — баз знаний. То ли в 80-х годах двадцатого столетия не до конца сформировалась общественная потребность в широком использовании таких систем, то ли недостаточными были мощности компьютеров, то ли не доработаны были теоретические и алгоритмические основы таких систем, но бум их популярности в конце 80-х годов закончился. За прошедшее с тех пор время наряду с бурным технологическим процессом (до сих пор не опровергнут закон

Мура) сложилось понимание того, что для решения проблемы информационного хаоса больше всего подходят технологии, порожденные некогда таким направлением, как контент-анализ, и сегодня получившие названия Data Mining и Text Mining. В настоящее время существуют достаточно развитые системы, реализующие эти направления. Практически все самые известные производители программного обеспечения предлагают на рынке системы глубинного анализа данных и текстов (у компании Oracle — это Oracle Text, у IBM — Intelligent Miner for Text, у SAS — Text Miner).

Следует отметить, что большая часть информационного потока — это неструктурированная текстовая информация, в то время как значительная часть электронной информации, порожденной путем использования современных СУБД, — это численные фактографические данные. Если обработка таких данных позволяет использовать уже отработанные методы и погружать потоки данных в СУБД, то задача анализа текстовой информации открывает широкое поле для применения новейших методик и технологий, таких как XML, лингвистические, эмпирические, статистические подходы. В настоящее время уже определено несколько задач, стоящих перед технологией Text Mining, — это автоматическая классификация, кластеризация, выявление смысловых взаимосвязей отдельных фрагментов и понятий, выраженных в тексте, а также составление осмысленных рефератов, резюмирующих знания, содержащиеся в текстовых массивах больших объемов. Возможно, эти технологические подходы в случае массового применения смогут облегчить ориентацию человека в постоянно расширяемом информационном поле, позволят ему адекватнее реагировать на происходящие события, уверенно принимать важные решения на основе концентрации знаний.

Развитие вычислительной техники и компьютерных сетей способствовало появлению систем, назначение которых — поиск в массивах полнотекстовых документов. К таким документам можно отнести, например, статьи, нормативные акты, реферативные описания, тексты брошюр, диссертаций, монографий. До определенного времени полнотекстовые информационно-поисковые системы использовались преимущественно специалистами, круг которых был не очень широк, — архивные работники, сотрудники библиотек, ученые, аналитики.

Появление и развитие сети Internet в корне изменило ситуацию. Сегодня информационные ресурсы Сети составляют около десяти миллиардов документов (Web-страниц), к которым возможен свободный доступ любого пользователя. Естественно, чтобы найти необходимую информацию в этой крупнейшей полнотекстовой базе данных, необходимо использовать очень мощные поисковые средства, которые в зачаточном состоянии уже существуют, развиваются и конкурируют друг с другом на рынке информационных технологий.

Сегодня миллионам пользователей Internet известны такие системы, как Google, Yahoo, AllTheWeb, AltaVista, каждая из которых охватывает несколько миллиардов Web-документов. Мы стали свидетелями “информационного взрыва”, в результате которого менее чем за 10 лет мало кому известная технология полнотекстового поиска стала повседневным инструментом миллионов людей.

В связи с этим первая глава книги — “New Media” — посвящена Internet и ее информационному подпространству World Wide Web. В этой главе описывается топология этого подпространства, а также средства навигации в нем и эволюция этих средств — от простейших наборов ссылок и каталогов до многофункциональных порталов.

Вторая глава посвящена поисковым системам, процессу поиска информации и его отдельным звеньям, а также включает трактовки таких фундаментальных понятий информационного поиска, как полнота и релевантность. Кроме того, эта глава содержит информацию о практической стороне использования процедур поиска, особенностях формирования запросов к различным информационно-поисковым системам с использованием слов, словоформ, фрагментов текстов, а также о поиске с учетом структуры документов, морфологии, подобия.

Третья глава охватывает вопросы ориентации в новостной информации, представленной в Сети. Для такого поиска используется специальный класс информационно-поисковых систем — системы мониторинга контента Internet, на основе которых строятся современные службы синдикации новостей.

Вопросам современного унифицированного представления информации в перспективном формате гипертекстовой разметки XML, а также технологическим решениям, построенным на основе идеологии XML, посвящена четвертая глава "XML — язык разметки и модель данных".

Технологиям выявления знаний в текстовых массивах с использованием как классических, так и новых, интеллектуальных подходов к анализу информации посвящена пятая глава "Технология Text Mining".

Шестая глава посвящена очень популярному сегодня направлению использования технологии Text Mining — конкурентной разведке, которая заключается в сборе и аналитической обработке информации, необходимой для принятия оптимальных управленческих решений. Очень важно, что при этом конкурентная разведка выполняется строго в рамках правовых норм.

Седьмая, заключительная, глава книги содержит обзор общих закономерностей, присущих информационным системам, в частности таких, как правило Парето, законы Зипфа и Брэдфорда и так далее, что должно дать читателю некоторое обобщенное представление о тенденциях и подходах, обсуждаемых в книге.

Дмитрий Ландэ, сентябрь 2004 года