ДОКУМЕНТАЛЬНЫЕ ИСТОЧНИКИ ИНФОРМАЦИИ

УДК 002.2

С. М. Брайчевский, Д. В. Ландэ

Современные информационные потоки: актуальная проблематика

Обсуждается современное состояние информационных технологий, связанное с быстрым ростом объемов информации и темпов ее распространения. Показано, что основные сложности вызваны не уровнем программного или аппаратного обеспечения, а специфическими особенностями предметной области. В рамках концепции информационных потоков проведен анализ актуальных проблем доступа к данным, структуризации информационного пространства и его связи с семантическим пространством, извлечения знаний из текстов, а также некоторых направлений решения этих проблем. Предлагается новый подход к обеспечению генерации, распространения и обработки данных, условно названный "сетевой навигацией". Обоснована перспективность исследования современных информационных потоков для специалистов в различных областях, например, в плане аналогового моделирования статистических процессов, в том числе сложных нелинейных систем с элементами самоорганизации.

ВВЕДЕНИЕ

Бурное развитие информационных технологий, в частности, сети Интернет, в последнее время породило ряд специфических проблем, связанных, в первую очередь, с быстрым ростом объемов данных, подлежащих хранению и обработке.

В начале существования World-Wide Web небольшое количество веб-сайтов публиковало информацию отдельных авторов для относительно большого количества посетителей. Сегодня ситуация резко изменилась. Сами посетители веб-сайтов активно участвуют в создании контента, что привело к резкому росту объема и динамики информационного пространства.

Сегодня в Интернете уже существует доступная для экспериментов информационная база такого объема, который ранее трудно было представить. Более того, объемы этой базы превышают на несколько порядков все то, что было доступно десятилетие назад. В августе 2005 г. компания Yahoo объявила о том, что проиндексировала около 20 млрд. документов. Прошлогоднее достижение компании Google составляло менее 10 млрд. документов, т. е. за один год количество открытой, доступной простому пользователю, информации из Интернета удвоилось. По данным службы Web Server Survey в августе 2005 г. количество веб-сайтов превысило 72 миллиона (рис. 1). Таким образом, приведенные данные подтверждают экспоненциальный характер роста информации.

Этот рост сопровождается рядом таких проблем [1], как:

- непропорциональный рост уровня информационного шума;
- засилье паразитной информации (невостребованной, получаемой в качестве несанкционированных "приложений");
- слабая структурированность информации;
- многократное дублирование информации.

Традиционному вебу к тому же присущи такие недостатки, как обилие "информационного мусора", невозможность гарантирования целостности документов, практическое отсутствие возможности смыслового поиска, ограниченность доступа к "скрытому" вебу.

Над решением названных проблем работают многочисленные коллективы ученых и специалистов во всем мире, в частности, консорциум W3C, где реализуется концепция Семантического веба [2]. Наряду с этой концепцией, революционный прорыв обещает дать более общий подход, а именно веб-2 (http:/www.web2con.com/), или, как его называют, "веб второго поколения", который предполагает реализацию концепции семантического веба, включая многоуровневую поддержку метаданных, новые подходы к дизайну и соответствующему инструментарию, технологию глубинного анализа текстов (Text Mining), а также идеологию веб-сервисов, базируясь при этом на информационных ресурсах, накопленных в WWW первого поколения.

Сегодня в связи с развитием информационных ресурсов сети Интернет документальное информационное пространство развилось до такого уровня, который требует новых подходов. Рост объемов информации и скорости ее распределения фактически породил понятие информационных потоков [3]. Вместе с тем, математический аппарат и инструментальные средства уже не всегда способны адекватно отражать ситуацию, речь идет не столько об анализе конечных массивов документов, сколько о навигации в документальных информационных потоках.

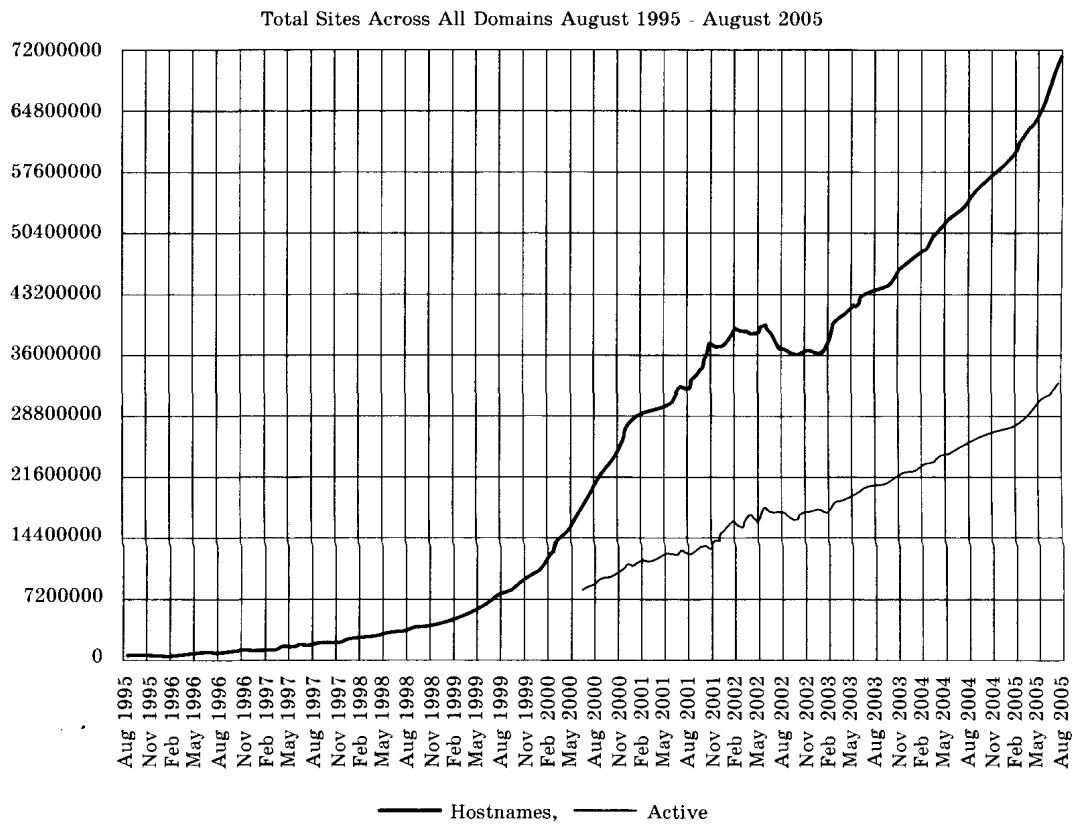


Рис. 1. Динамика роста количества веб-сайтов

ФИЛОСОФСКИЕ КАТЕГОРИИ "ИНФОРМАЦИЯ" И "ЗНАНИЯ"

Сегодня есть основания полагать, что определенного переосмысления требует само понятие информации, в частности его взаимосвязь с понятием "знания", ставшим особенно популярным в последние годы. Часто употреблявшийся ранее в теории искусственного интеллекта и впоследствии основательно забытый термин "преобразование информации в знания", похоже, вновь начинает вызывать интерес.

Этому в значительной мере способствовали чисто прикладные успехи в машинной обработке потоков данных, содержащих документы, не только составленные на разных языках, но и относящихся к различным социокультурным контекстам. Ясно, что в таком случае обработка потока данных (т. е. информации в чистом виде), какой бы она ни была, не предполагает активного использования содержания документов. Теоретическое осмысление подобных ситуаций наводит на мысль о том, что собственно "знания" представляют собой некую надстройку над информационными потоками, определяемую в конечном счете наличием устойчивых связей между определенными информационными элементами. Подчеркнем, что эти связи сами по себе в информационных потоках не содержатся и являются в этом смысле внешним по отношению к информации фактором.

Практика показывает, что информация может вполне успешно обрабатываться вне зависимости от того, какой смысл в нее заложен. В связи с этим вновь возник интерес к подходам, основанным на понимании информации как меры упорядоченности некоторой системы и, соответственно, к статистическим методам ее обработки.

Некоторые ученые и ведущие участники информационного рынка (в частности, компания Autonomy) возвращаются к истокам теории информации, понятиям энтропии, теории Шеннона, уравнениям Больцмана и др.

При этом оказалось, что многие задачи, возникающие при работе с информационными потоками, имеют немало общего с задачами статистической физики и гидродинамики и могут решаться одними и теми же методами. Это обстоятельство открывает широкие перспективы применению мощного аппарата современной физики к решению теоретико-информационных задач.

С другой стороны, признание того, что извлечение из информационных потоков знаний в обычном смысле слова является самостоятельной проблемой, которая должна решаться методами, требующими отдельной разработки, и несомненно будет способствовать развитию этих методов, так же, как и соответствующих инструментальных средств.

Если информация может обрабатываться независимо от содержательного аспекта, то обратное не верно. В любом случае именно информация является своего рода "субстратом знаний".

Поэтому теория информации, которая ранее находила свое основное реальное применение в области техники передачи информации, сейчас становится полезной и для анализа смысловых текстовых потоков. Энтропия с помощью осмысленного анализа уменьшается весьма постепенно, но чем этот анализ комплекснее, тем заметнее переход от хаоса к порядку.

ИНФОРМАЦИОННОЕ И СЕМАНТИЧЕСКОЕ ПРОСТРАНСТВО

Проблема "знаний", скорее всего, никогда не будет сведена к какому-либо комплексу задач, которые можно было бы окончательно решить чисто технологическим путем. Напротив, она, видимо,

потребует серьезных исследований в различных направлениях, в том числе и на достаточно высоком теоретическом уровне.

Одним из центральных вопросов в этом плане, на наш взгляд, является отношение информационного и семантического пространства, чему, как правило, уделяется неоправданно мало внимания. В литературе их часто даже отождествляют без всякого на то основания. То, что эти две категории никоим образом не тождественны, с очевидностью вытекает из различия их природы: информационное пространство образуют данные, физически записанные на тех или иных носителях, тогда как семантическое пространство порождают комплексы абстрактных понятий, связанных с субъективными оценками, даваемыми человеком. Наиболее естественным представляется определить сетевое семантическое пространство как множество единиц смысла, актуальных в данном социокультурном контексте и представленных в сети. Под единицей смысла мы, как обычно, понимаем элементарную категорию, позволяющую нам строить субъективные оценочные суждения о вещах и процессах, относящихся к окружающему нас миру.

В реальной жизни между ними, безусловно, существует вполне определенная связь, но отыскание этой связи, по-видимому, представляет собой весьма нетривиальную задачу.

Глубину и важность задач понимания соотношения информационного и семантического пространств проиллюстрируем на примере автоматического реферирования текстового массива, содержащего документы, составленные на разных языках. Возможен ли алгоритм, позволяющий выделить из произвольного документа информационно значимые фрагменты, "не зная" языка, на котором этот документ создан? Разумеется, речь не идет об идентификации языка документа с последующим подключением соответствующих словарей, наборов грамматик и т. п.

Оказывается, такой алгоритм возможен, если только входной поток удовлетворяет законам Ципфа, т. е. создан человеком. Более того, он успешно реализован авторами в системе InfoStream [1]. И это порождает вполне еретический вопрос: в какой мере понятие "информация" связано с понятием "смысл", и связаны ли эти понятия вообще? По крайней мере, в общепринятом понимании.

Действительно, приведенный пример показывает, что определенное количество информации может быть передано и надлежащим образом обработано (с чем конечный потребитель, прочитав реферат и сравнив его с оригиналом, согласится), но при этом в процессе обработки смысл, возможно содержащийся в ней, никак не учитывался. Более того, мы можем даже не знать, имеют ли вообще какое-либо значение в семантическом отношении последовательности символов, составляющие документ ("Глокая куздра штеко будланула бокра...").

Самое интересное при этом заключается в интерпретации полученных результатов. Без привлечения методов искусственного интеллекта, объемных семантических нормализаторов, даже экспертов как таковых, с использованием только частотных методов могут быть получены содержательные, семантически наполненные результаты. Возникает ощущение, что для полноценной работы с информацией вполне достаточно структурно-лингвистического уровня.

При желании можно было бы возразить, что поскольку смысл сопряжен с понятиями, т. е. знаками вещей, а не с их знаками (т. е. знаками знаков — словами), он, вообще, говоря, от языка не зависит, и этот факт позволяет осуществлять переводы текстов. Фокус, однако, в том, что алгоритм, о котором идет речь, переводом (как бы мы его ни понимали) как раз и не занимается: он просто "не обращает внимания" на содержание документа.

В качестве других примеров можно привести автоматическое выявление взаимной связи понятий, автоматическую кластеризацию связей для выявления наиболее важных из них, автоматическое выявление "окраски" взаимосвязей, в простейшем случае — определение принадлежностей взаимосвязей к положительным (группирующим) или отрицательным (антагонистическим).

Вместе с тем, потребитель все же хочет в конечном счете получить нечто осмысленное. Поэтому полное игнорирование семантических аспектов в информационных технологиях было бы ошибочным. Видимо, оптимальный путь состоит в том, чтобы более адекватно оценить функциональную роль и значение семантического уровня информационных процессов. Видимо поэтому, сегодня все чаще возникает вопрос о природе связи информационного и семантического пространств.

Для соотнесения элементов информационного и семантического пространств необходим некий промежуточный модельный уровень обработки текстовых данных. При этом должны быть определены "правила чтения", с помощью которых формальная система (набор структурных элементов текста) преобразуется в систему содержательную (осмысленное сообщение). Более того, эти правила должны быть встроены в некую компьютерную программу. И здесь возникает серьезная сложность. В реальной жизни такие правила никогда не формализуются. Человек постигает их годами, активно действуя в определенном социокультурном контексте, постоянно общаясь с другими людьми. Причем различные контексты порождают различные "правила чтения", которые, к тому же, изменчивы во времени.

В наше время не существует единого способа научить таким правилам машину, а без этого, в свою очередь, невозможно добиться того, чтобы она, обрабатывая текст, учитывала его содержательный аспект предсказуемым образом.

Таким образом, с информационным пространством оказывается сопряжено не только семантическое пространство, которое может быть доступно нашему интеллекту, но и пространство формальных "правил чтения", позволяющих производить заданный набор операций.

"Обратной стороной медали" является тот несомненный факт, что информационное пространство в конечном счете порождается семантическим. Действительно, возникновение информационных потоков можно представить себе как генерацию и движение наборов данных, ассоциированных с определенным сообщением, понимаемым как некоторый смысловой блок. Конечно, одному сообщению может соответствовать произвольное число отдельных наборов данных (о важном международном событии напишут все медийные средства).

Таким образом, характеристики информационного пространства изначально определяются структурой сетевого семантического пространства, причем здесь мы имеем право говорить о структуре, поскольку сообщения отражают события реального мира, который, вероятно, все же в какой-то мере упорядочен.

ПРОБЛЕМЫ ПОИСКА ИНФОРМАЦИИ

Эпиграфом к данной теме могли бы послужить слова персонажа кинофильма "Уолл-стрит": скажи мне то, чего я не знаю!

Попытки технологического развития в рамках современной теории информационного поиска сегодня очень часто не улучшают, а ухудшают ситуацию. Например, совершенствование технических аспектов информационно-поисковых систем приводит лишь к увеличению объемов релевантных наборов, которые часто не пригодны к употреблению.

Современные технологии позволяют осуществлять невероятно изощренные операции над данными, но чем эффективнее они применяются, тем менее "съедобным" оказывается результат.

Надежды, возлагавшиеся в свое время на идею последовательного уточнения поиска ("искать в найденном", "показать подобное" и т. п.), не оправдались по двум причинам. Во-первых, интересующий потребителя документ может просто не оказаться в первичной выборке, в силу чего последующие итерации теряют смысл, а во-вторых, составление уточняющего запроса, качественно отличающегося от исходного, представляет собой отнюдь не простую задачу, прямо скажем, непосильную для рядового пользователя.

Мы не ищем в сетях то, что и так знаем — в этом нет никакого смысла. Нам нужно что-то, чего мы не знаем, и мы лишь пытаемся объяснить машине, в каком сегменте информационного пространства это "что-то", по нашему мнению, должно находиться. Указание набора слов, с помощью которых этот сегмент можно локализовать, оказывается отнюдь не самым совершенным способом достижения поставленной цели.

Очевидно, следует признать, что изначальная парадигма поисковых систем, сформированная десятки лет тому назад, уже не отвечает реальной ситуации. Таким образом возникает задача поиска новых принципов, в рамках которых оказалось бы возможным проектирование качественно новых систем обработки больших и динамичных объемов данных.

Поисковые машины следующих поколений должны будут лучше классифицировать информацию и нагляднее представлять ее. В будущем поиск не должен ограничиваться лишь обработкой введенных ключевых слов.

Имеет смысл реализация перехода к концепции навигации в информационных потоках как к распределенному во времени интерактивному процессу локализации отдельных семантических секторов в общем информационном потоке.

Системы должны будут отслеживать интересы пользователей, делая поиск более целенаправленным. Новые поисковые машины будут находить опубликованные в сети текстовые, аудио- и видеоматериалы, которые в настоящее время недоступны.

В последнее время получили распространение адаптивные интерфейсы уточнения запросов [3], чаще всего реализуемые методами кластерного анализа. Появилось такое понятие, как метод "папок поиска" (Custom Search Folders), который не связывается с определенным алгоритмом, а представляет собой множество подходов, общее у которых — попытка сгруппировать данные и представить кластеры в удобном для пользователей виде.

К подобным механизмам можно отнести, например, австралийский поисковый сервер Mooter (http://www.mooter.com), на котором применяется визуальный подход к предоставлению результатов поиска по обрабатываемым запросам путем группировки результатов первичного поиска по категориям. Другой поисковый сервер iBoogie (http:www.iboogie.com/) также группирует результаты поиска, но отображает их в виде, близком к экрану проводника Windows.

Недавно разработчики Google представили свои наработки и планы по кластеризации найденных документов. Демо-версия этой системы позволяет выделять из документов названия компаний, которые являются основными критериями кластеризации.

Одним из наиболее интересных решений следует считать метод так называемых информационных портретов, использующихся для уточнения запросов. Однако тут уточнение осуществляется за счет добавления не произвольных терминов, придумываемых пользователем, а определенного их набора, формируемого машиной в процессе статистической обработки доступного массива данных. Иными словами, пользователю предлагается то, что реально существует (он может обнаружить в списке термин, вполне отвечающий его потребностям, но который сам он не смог бы придумать). Поэтому, возможно, правильнее было говорить не об уточнении поиска, а о его сужении.

На рис. 2 представлен пример сужающего списка терминов, используемого нами при обработке потоков новостной информации в рамках технологии оригинальной InfoStream. В частности, в адаптивном интерфейсе системы существенно облегчен множественный выбор источников информации, соответствующих заданному запросу. Предусмотрен и такой "экзотический" параметр, как уровень насыщенности документов цифровой информацией, что полезно, например, при поиске аналитических документов, ценовых таблиц, рейтингов и т. п.

Воплощением идеи коллективной работы в Интернет, входящей в концепцию информационно-поисковых систем нового поколения, сегодня стала система с "хвостовыми данными" Snap, обеспечивающая не только поиск веб-страниц по ключевым словам, но и предоставляющая дополнительную информацию, близкую интересам пользователей. Например, к результату поиска по изготовителям цифровых камер добавляется сравнительная таблица моделей, которые ранее были затребованы другими пользователями системы. Данная поисковая система является предвестником такого этапа развития WWW, на котором в ней будут активно использоваться результаты работы всего сообщества пользователей.

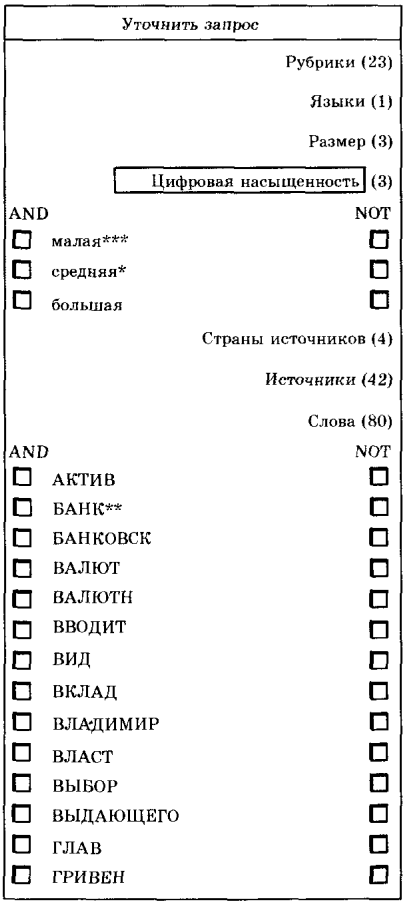


Рис. 2. Информационный альбом системы InfoStream

Таким образом, резюмируя приведенные выше рассуждения, выскажем предположение о том, что современные информационные технологии готовы к пересмотру принципов обеспечения доступа к сетевым данным, который условно можно назвать переходом от информационного поиска к сетевой навигации.

РЕЛЕВАНТНОСТЬ И ПЕРТИНЕНТНОСТЬ

Вероятно возможно вычленить центральную проблему современных информационных потоков — она состоит в качественном различии понятий "релевантность" и "пертинентность". Сам факт наличия этих двух терминов говорит о том, что различие было известно всегда, но в условиях ограниченных объемов данных им можно было пренебречь, так как потребитель, в явном виде просмотрев всю релевантную выборку, мог отобрать то, что ему нужно. Сегодня же, когда часто это становится невозможным, несовпадение релевантности с пертинентностью выступает на первый план. Действительно, если из 10 тысяч предъявленных ИПС документов *все* являются пертинентными, то потребитель будет удовлетворен, по крайней мере в первом приближении, прочитав *любое* их количество. Остальные он может просто проигнорировать без особого ущерба для достижения поставленной цели. В некоторых областях отмеченная закономерность эффективно используется. Например, службы синдикации новостей обслуживают своих клиентов, при том, что количество охватываемых источников информации практически у любой из них в настоящее время не превосходит 10 тысяч. При этом следует отметить, что проблему полноты новостной информации такой подход позволил решить, оставив, однако нерешенной проблему формирования достаточного для пользователя объема информации.

Существующие ИПС изначально проектировались для обеспечения именно релевантности выборки по отношению к формальным запросам, и в этом их главная слабость в современных условиях. Низкий, а точнее говоря, неконтролируемый уровень пертинентности выборки с высоким уровнем ее релевантности порождает различные ситуации, допускающие более или менее общую типизацию.

[Наиболее простой и очевидный случай: по запросу "президент" можно получить кроме необходимых новостей рекламу отеля "Президент", прайс-листы сигарет "Президент" и т. п. Теоретически, с таким информационным мусором можно бороться, составляя изощренные запросы из 200–300 поисковых терминов с активным использованием контекстной близости и операций отрицания, но это сложная работа, требующая времени, определенной подготовки и практического опыта. Во всяком случае, у обычного пользователя есть шанс получить желаемое, прибегнув к помощи профессионалов.

Предположим, пользователя интересуют специальные работы по методам кодирования текстовой информации. Он составляет соответствующий запрос и получает набор документов, которые действительно посвящены этой теме. Но ему предъявляются классические учебники теории связи, тогда как требуется получить последние публикации с оригинальными результатами. Здесь уже расширение запроса, скорее всего, не поможет, поскольку его обработка, какой бы сложной она ни была, предполагает использование того, что содержится в тексте документа в явном виде и может быть реализовано лингвистическими средствами.

Далее, пусть пользователь действительно получил ссылку на обзор по теме, содержащий то, что ему нужно. Но неприятность при этом заключается в том, что оказывается, что именно этот обзор уже лежит у него на столе, и, значит, нет нужды искать его в базах данных, а другие обзоры, возможно, находятся где-то в конце списка ссылок, но этого он никогда не узнает. И справиться с такой ситуацией намного сложнее, чем с первыми двумя, потому что факт наличия у пользователей некоторых данных никак не отражен в самих данных.]

ПРОБЛЕМА СТРУКТУРИЗАЦИИ ИНФОРМАЦИИ

Вряд ли есть смысл отдельно говорить о том, что сетевое информационное пространство в принципе структурируемо достаточно слабо. Более того, эволюция как сети в целом, так и отдельных ее сегментов может служить некоторым примером стохастического процесса. Возможно, именно это обстоятельство и является главной причиной низкой эффективности организации быстрого прямого доступа к информационным единицам, о которых мы часто даже не знаем, существуют ли они вообще в данный момент времени.

Сказанное не означает, что сетевое пространство является хаотическим и может быть полностью описано в терминах шума. На самом деле, оно

содержит в себе элементы упорядоченности (назовем их кластерами), в известной мере аналогичные доменам ферромагнетиков. Но их много, и каждый из них обладает собственной динамикой развития, слабо коррелирующей с другими такими динамиками. С другой стороны, кластеры могут интенсивно взаимодействовать друг с другом, порождая своего рода "отраженные волны" и формируя тем самым разнообразные обратные связи, как положительные, так и отрицательные.

Но, во-первых, сами кластеры не всегда являются устойчивыми во времени — они возникают, исчезают, меняют свои контуры, мигрируют и т. п. и, во-вторых, взаимодействие между ними носит стохастический характер.

Первый реальный шаг к решению проблемы структуризации сетевого информационного пространства, очевидно, должен состоять в формировании некоего порожденного пространства, обладающего достаточным уровнем упорядоченности и в разумном приближении адекватного исходному. Таким образом, может быть поставлена задача, понимаемая как некоторое, вообще говоря, неоднозначное, отображение неупорядоченного множества составляющих элементов сетевого информационного пространства на упорядоченное множество их образов, обладающее требуемой (например, иерархической) организацией.

Тогда поиск в широком смысле слова может производиться на структурированном множестве образов информационных единиц, а предъявление его результатов должно включать в себя восстановление оригиналов.

Конечно, приходится считаться с возможностью утраты части информации, но ведь и в стандартной реализации невозможно добиться одновременно релевантности результата и ее разумной полноты. Поэтому, так или иначе, мы можем лишь говорить о некотором ожидании получения требуемого, связанного с вероятностью его обнаружения. По крайней мере, для определенного класса задач такой подход, в числе прочего, может решить до сих пор открытую проблему теории поиска — проблему информационного дублирования. Именно при построении пространства образов могут формироваться цепочки более или менее информационно-подобных элементов, отображаемые затем на один и тот же образ. Естественно, в процессе построения пространства образов, они могут снабжаться наборами метаданных.

Первой попыткой практического решения названной проблемы являются рубрицированные каталоги веб-сайтов, однако ее следует считать ограниченной по двум причинам: во-первых, как правило, классифицируются только сайты, а не входящие в их состав документы, а во-вторых, используется стандартный (и практически не зависящий от времени) набор предопределенных рубрик. Суть проблемы даже не в том, что рубрик слишком мало для полноценной структуризации сетевого информационного пространства, а в том, что они отражают не его реальные свойства, а субъективные представления потребителей о структуре предметной области. Например, внешняя статическая рубрикация не в состоянии локализовать реально существующие в данный конкретный момент кластеры.

По мнению авторов, структуризация сетевого информационного пространства неизбежно должна предполагать постоянное (фоновое) сканирование информационных потоков и создание их виртуального образа, предназначенного для практического использования. В идеале, этим могли бы заниматься специализированные службы, предоставляя результаты своей деятельности в распоряжение поисковых систем.

Одним из наиболее естественных путей решения подобных проблем нам представляется перенесение центра тяжести с наборов данных, в которых следует вести поиск, на ассоциируемые с ними наборы метаданных, содержащих широкий спектр внешних характеристик, по которым потребитель может достаточно просто сформировать своего рода "словесный портрет" требуемых документов. Ядро формируемого пользователем поискового предписания должно представлять собой набор формальных параметров, указывающих на выбор той или иной категории из содержащихся в метаданных. Традиционный же запрос, включающий в себя ряд поисковых терминов, может играть здесь вспомогательную роль для сужения объема (теперь уже пертинентной) выборки.

Наборы метаданных могут, разумеется, создаваться специальными программными комплексами, входящими в состав поисковых систем, путем автоматической обработки сканируемой информации. Однако несравненно большей эффективности можно было бы достичь, организовав хранение метаданных (пусть даже "сырых") непосредственно в информационных документах.

Для этих целей как нельзя лучше подходят XML-технологии, получившие в последнее время широкое распространение. Например, в состав вебсайта мог бы входить XML-документ, содержащий некое унифицированное (в идеале — стандартизированное) описание структуры и основных предопределенных характеристик этого сайта, предназначенных для построения различного рода классификаторов.

Легко видеть, что даже в простейшем случае подключения набора классификаторов мы получаем значительный выигрыш в объеме результатов поиска. Пусть мы имеем три классификатора, каждый из которых содержит десять категорий. В случае равномерного распределения документов по категориям получим фактор 10^{-3}, т. е. вместо 10 000 документов потребитель получит их всего 10.

Разумеется, приведенные рассуждения касаются не только поиска в чистом виде, но и сопряженных задач таких, как, скажем, избирательное распространение информации.

РАНЖИРОВАНИЕ ИНФОРМАЦИОННЫХ ПОТОКОВ

В свое время была высказана идея, что информационный поиск в классическом понимании вообще не нужен, если имеется достаточно эффективная процедура сортировки данных по набору параметров, количественно выражающих информационную потребность пользователя. В этом случае система просто последовательно выдает все документы, содержащиеся в базе, но в начале списка стоят именно те, которые наиболее соответствуют поставленной задаче.

В чистом виде эта идея, вероятно, слишком радикальна, но доля истины в ней, безусловно, присутствует. В самом деле, традиционная процедура информационного поиска, в конечном счете, сводится к определению для каждого документа релевантности относительно предложенного запроса.

Известно, что для все еще популярной и в настоящее время модели поиска по инвертированным словарям, релевантность может иметь только два значения: 0 или 1, т. е. классический поиск всего лишь выделяет из генеральной совокупности выборку документов, являющихся в этом смысле равноценными и равнозначными.

Поэтому, если мы желаем приблизить результаты поиска к отражению реальной информационной ситуации, то он должен быть дополнен некоторой процедурой, позволяющей строить те или иные распределения по параметрам, допускающим субъективную оценочную интерпретацию, с последующей сортировкой результатов.

Перспективным путем реализации такой схемы, на наш взгляд, могло бы быть использование многопрофильных шкал, сформированных на основании некоторых метаданных.

Новое звучание должен получить и кластерный анализ. Следует заметить, что сама информация генерируется бессистемно. Без специальных акций и программ ее порождает большое количество источников. Кластерный анализ информационных потоков призван обеспечить постоянный и надежный процесс систематизации. Проблема в том, что большинство известных методов ориентировано на кластеризацию статических объектов, в то время как информационное пространство представляет собой динамическую систему [4].

Однако эта проблема, как ни парадоксально, открывает новые возможности, качественно отличающиеся от возможностей статической кластеризации, а именно — мы можем учитывать временные зависимости основных параметров информационных потоков. Например, исключительно важным может оказаться изучение временной устойчивости статистических характеристик потоковых данных.

*В рамках традиционной пространственно-векторной модели, на базе которой применяются алгоритмы кластерного анализа, вес отдельных термов (слов и словосочетаний) вычисляется как TF * IDF. Напомним, что TF — это локальная частота терма (Term Frequence), а IDF — величина, обратная частоте встречаемости во всем потоке документов, содержащих данный терм (Inverse Document Frequency). В то время как локальная частота терма в документе говорит о значимости терма в пределах документа, обратная частота встречаемости свидетельствует об уникальности терма во всем потоке документов. Поэтому произведение этих величин — достаточно удачный критерий определения веса терма в стационарном массиве документов. В случае обработки информационных потоков мы можем дополнительно исследовать изменение во времени отношения этих частот. Действительно, значимость документа в потоке в значитель-*

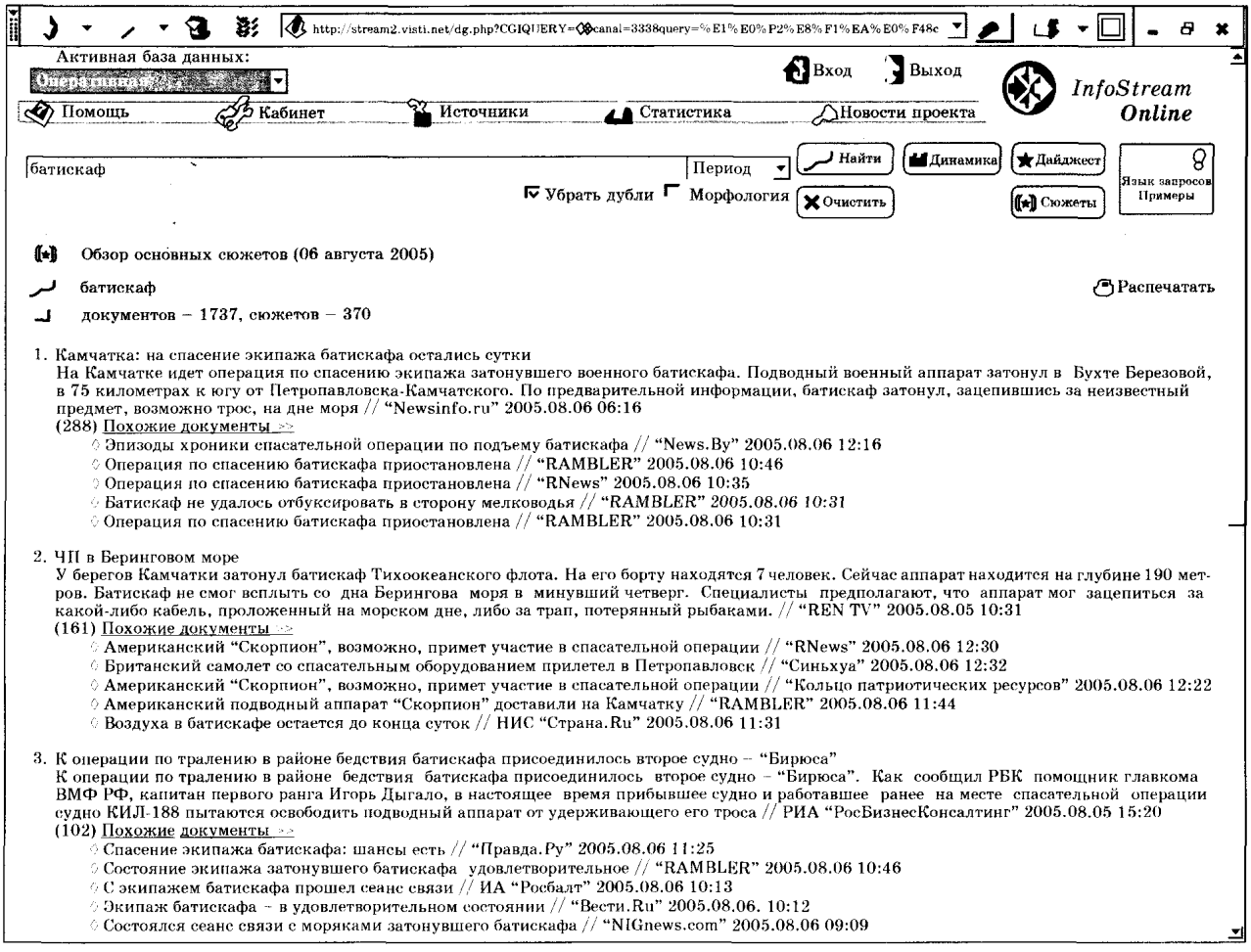


Рис. 3. Основные сюжетные цепочки по теме "Батискаф"

ной мере определяется изменением во времени относительного числа документов аналогичного содержания, поскольку информационные потоки так или иначе отражают динамику развития реальных ситуаций. Один из подходов к кластеризации потоков документов, по мнению авторов, должен заключаться в использовании поправочных множителей, зависящих от времени. Например, можно определять вес документа как сумму элементов типа $TF^* IDF^* e^{-\alpha t}$, *где* α *— некоторая константа,* t *— интервал времени, прошедший с момента появления документа в информационном потоке (значение* α *— это коэффициент полураспада ранга документа, т. е. предполагается, что* $e^{-\alpha t} = 1/2$, *где* t *— экспертно определяемый период времени, в течение которого документ ввиду устаревания теряет свою актуальность наполовину). Например, если предположить, что за сутки документ теряет половину своей актуальности, то имеем:* $e^{-\alpha^* 24} = 1/2$, *и, соответственно,* α*=0,025.*

Названный подход, позволил, в частности, авторам в рамках технологии InfoStream реализовать построение основных сюжетных цепочек на основе тематической выдачи информационно-поисковой системы (Рис. 3).

Следует отметить, что задача кластерного анализа очень важна, так как динамически меняющийся набор кластеров задает некую концептуальную сеть, в терминах которой анализируется информационный поток. При этом часто используется и человеческий фактор, выраженный в экспертных оценках, которые выступают в качестве обратного контура обучаемых систем. Выявление кластера предполагает также его описание. В семантическом подходе к описанию, при всей сложности и многообразии характеристик описания, присутствуют их количественные оценки. Кроме того, современные информационные потоки практически содержат в себе весь словарь современного языка, мало того — "готовые" специализированные словари: частотный, инверсный и прочие.

ИЗВЛЕЧЕНИЕ ЗНАНИЙ ИЗ ТЕКСТОВ

Поисковые технологии должны стать более эффективными за счет технологий, объединяющих поиск и глубинный анализ текстов (Text Mining), нахождение в текстах аномалий и трендов. Только в этом случае поиск информации станет связанным с ее осмыслением.

В настоящее время выделяют четыре основных вида приложений технологий Text Mining:

- Классификация текстов благодаря выявлению статистических корреляций для формулирования правила размещения документов в предопределенные категории.
- Кластеризация, базирующаяся на выявлении латентных признаков документов, применяющая лингвистические и математические методы без использования предопределенных категорий, что может дать эффективный охват больших объемов данных.
- Построение семантических сетей на основе анализа документальных информационных потоков.
- Извлечение фактов из текстов.

Уже из одного этого перечня названий видно, что к собственно знаниям эти приложения имеют весьма отдаленное отношение. Их скорее можно рассматривать как некую промежуточную платформу, облегчающую дальнейшие манипуляции с данными. Например, в плане поставленной цели извлечение фактов из текста имеет смысл лишь в том случае, когда предвидится дальнейшее установление определенных отношений между ними — разрозненные факты, лишенные связей, претендовать на знания ни в коей мере не могут.

Можно назвать еще несколько задач технологии Text Mining, например, прогнозирование, которое состоит в том, чтобы предсказать по значениям одних признаков объекта значения остальных. Еще одна задача — нахождение исключений или аномалий, т. е. поиск объектов, которые своими характеристиками сильно выделяются из общей массы. Для этого сначала выясняются средние параметры объектов, а потом исследуются те объекты, параметры которых наиболее сильно отличаются от средних значений. Сегодня подобный анализ часто проводится после классификации для того, чтобы выяснить, насколько последняя была точна.

Здесь мы уже видим отчетливые, хотя и явно недостаточные элементы перехода от формальных систем к содержательным.

Несравненно ближе к решению общей проблемы извлечения знаний из текста стоит задача поиска скрытых связей отдельных признаков (дескрипторов, понятий). От предсказания эта задача отличается тем, что заранее неизвестно, по каким именно признакам реализуется взаимосвязь; цель именно в том и состоит, чтобы найти связи признаков. Решение этой задачи позволяет сокращать размерность пространства признаков, создавать обозримые классификаторы, пригодные для решения задач навигации в информационных потоках.

И наконец, для обработки и интерпретации результатов Text Mining большое значение имеет визуализация. Визуализация на основе систем Text Mining предполагается как средство представления контента всего потока документов, а также для реализации навигационного механизма, который может применяться при исследовании документов и их классов.

Рискнем предположить, что это — наиболее выдающееся достижение современных информационных технологий в данном направлении. Суть заключается в том, что эффективное представление потоковых данных в удобной для пользователя форме позволяет непосредственно задействовать человеческий интеллект, который, в конечном счете, намного быстрее приведет к поставленной цели, чем любая машина. Собственно, машина и должна предоставить пользователю в удобной форме то, с чем он в общем случае сможет справиться и сам.

Как бы там ни было, приходится признать, что проблема извлечения знаний из текста находится, видимо, на стадии осмысления и, вероятно, в ближайшее время начнет бурно развиваться.

СЕМАНТИЧЕСКИЙ ВЕБ

Одним из наиболее перспективных, по-видимому, следует считать наметившееся в последнее время направление, которое можно условно назвать

синтетическим. Его основополагающая идея заключается в попытке решать комплексы достаточно сложных задач, исходя из неких единых принципов организации генерирования, транспортировки и обработки данных. Главный акцент при этом ставится на согласованность параметров объектов обработки и ее инструментальных средств.

В качестве примера, на котором нам хотелось бы остановиться подробнее, приведем Семантический веб, который ее создатели считают абсолютно самодостаточным. Концепцию Семантического Web [2], которую на международной конференции XML-2000, прошедшей в 2000 г. в Вашингтоне, выдвинул Тим Бернерс-Ли, заключается в организации такого представления данных в сети, чтобы допускалась не только их визуализация, но и их эффективная автоматическая обработка программами разных производителей. Путем таких радикальных изменений концепции традиционного веба предполагается превращение его в систему семантического уровня. Семантический веб должен обеспечить "понимание" информации компьютерами, выделение ими наиболее подходящих по тем или иным критериям данных, и уже после этого — предоставление информации пользователям.

Семантический веб можно представить как симбиоз двух направлений, первое из которых охватывает языки представления данных. На сегодняшний день основными такими языками являются Расширяемый язык разметки XML (eXtensible Markup Languge) и Средства описания ресурсов RDF (Resource Description Framework). Существует также ряд других форматов, однако XML и RDF предоставляют больше возможностей, потому они обладают статусом рекомендаций W3C.

Второе, концептуальное направление несет в себе теоретическое представление о моделях предметных областей, которые в терминологии Семантического Web называются онтологиями. 10 февраля 2004 г. консорциумом W3C была утверждена и опубликована спецификация языка сетевых онтологий OWL (Web Ontology Language).

Таким образом, две ветви Семантического Web используют три ключевых языка (соответственно, технологий):

спецификация XML, позволяющая определить синтаксис и структуру документов;

механизм описания ресурсов RDF, обеспечивающий модель кодирования для значений, определенных в онтологии;

язык онтологий OWL, позволяющий определять понятия и отношения между ними.

Семантический веб использует также и другие языки, технологии и концепции, в частности, универсальные идентификаторы ресурсов, цифровые подписи, системы логического вывода и т. д.

Практическая реализация Семантического веба зависит от существования веб-страниц, содержащих метаданные, формирование которых не входит в стандартный процесс веб-разработки. Вряд ли удастся заставить авторов Web-страниц вручную индексировать свои ресурсы с помощью терминологических словарей, онтологий Семантического веба. Очевидно, что интегрировать существующие ресурсы WWW в Семантический веб можно только автоматически. Данная задача является очень сложной, требует подходов, близких к технологии глубинного анализа текстов.

В качестве достаточно успешного примера реализации такого подхода, можно привести технику и методологию австрийско-швейцарской группы разработчиков, предназначенную для создания семантически аннотированных веб-страниц. Технология WEESA (WEb Engineering for Semantic web Applications) позволяет осуществлять автоматическую генерацию метаданных в формате RDF для структурированного контента. Для генерации метаданных используется Java-программа, которая на основе содержания одного или нескольких атрибутов в качестве исходных данных возвращает стандартную триаду RDF ("объект — атрибут — значение"). По утверждению авторов технологии, они уже успешно применили технику WEESA для обработки веб-приложений на сайте Международного венского фестиваля. Там был магазин билетов, более 60-ти описаний различных мероприятий, а также архив за последние 52 года. Эксперимент показал, что WEESA хорошо подходит для разработки веб-приложений Семантической Сети.

В качестве одной из первых популярных реализаций элемента Семантического веба сегодня можно признать и RSS-технологии. Диалект языка XML — формат RSS (Really Simple Syndication, Rich Site Summary, RDF Site Summary), специально предназначенный для легкого и быстрого обмена содержанием веб-сайтов. Применение RSS обеспечивает согласованный способ резюмировать содержимое веб-сайтов, а кроме того, его применение позволяет администраторам сайтов новостей, блогов, форумов и других часто обновляемых веб-ресурсов, получать простой унифицированный метод подачи информации о происходящих событиях.

Сегодня RSS принято рассматривать и как формат, предназначенный для публикации и обеспечения экспорта новостей на новостных сайтах [5]. После того, как информация преобразована в формат RSS, программа, ориентированная на этот формат, может загружать сведения об обновлениях веб-сайтов и в зависимости от результата, выполнять определенные действия, например, автоматически обновлять список актуальных информационных сообщений.

Пользователи могут получить доступ к данным в формате RSS с помощью специальных программ, называемых RSS-агрегаторами. Программа-агрегатор позволяет группировать публикации из различных источников, обеспечивая возможность одновременно следить за появлением новостей на всех сайтах, не требуя посещения каждого сайта в отдельности. При этом, конечно же, не требуется загружать из Сети лишнюю информацию, относящуюся, например, к оформлению веб-страниц.

Программы-агрегаторы выполняют синтаксический разбор данных, представленных в формате RSS, после чего могут реализовывать любые действия по отношению к этим данным, к примеру, отображать их на выбранном веб-сайте.

Перспективность и популярность RSS как стандарта обусловлена прежде всего его доступностью и простотой. Сегодня практически все ведущие информационные сайты в мире используют RSS как инструмент оперативного представления своих обновлений.

Возможно, именно на этом пути достигнуты наиболее впечатляющие результаты.

ТОПОЛОГИЯ ИНФОРМАЦИОННОГО ПРОСТРАНСТВА

Существуют некоторые попытки изучения топологии информационного пространства, однако

четкой теории или даже адекватной модели в последние несколько лет предложено не было. Веб-пространство, являясь, пожалуй, самой динамичной частью информационного пространства, характеризуется большим количеством скрытых в нем неявных экспертных оценок, реализованных в виде гиперссылок. Еще в 1999 г. Андрей Брёдер (Andrei Bröder) и его соавторы из компаний Alta Vista, IBM и Compaq построили модель ресурсов и гиперсвязей Сети [6].

Исследования опровергли расхожее мнение, будто Интернет — это единое пространство. В рамках этой модели было обнаружено постоянное соотношение между отдельными частями веб-пространства: центральное ядро (~28% ресурсов), "отправные веб-страницы" (~22%), "оконечные веб-страницы" (~22%), соединяющие их "мысы" и "перешейки" (~22%), острова (~6%). Четыре первых множества — более 90% веб-ресурсов, топологически относящихся к одной компоненте связности, и обусловили название модели — Bow Tie ("галстук-бабочка").

Знание топологии информационного пространства позволяет реализовать концепцию сетевой навигации (как прямой, следуя гиперссылкам, так и обратной), которая основана на предоставлении потребителю именно того, что ему нужно на самом деле — возможности наглядно ориентироваться в сети, перемещаться в ней и извлекать из нее то, что он отыщет в процессе работы. Причем в идеале вполне может оказаться, что найденное выглядит совсем не так, как предполагалось вначале.

Прообразом сетевой навигации может служить стихийный серфинг в Интернете. Некоторые пользователи догадались, что переходя с сайта на сайт, можно обнаружить такие материалы, которые целенаправленно найти не в состоянии все поисковые системы, вместе взятые.

В последнее время получила большую популярность теория фракталов и хаоса, которая находит свои приложения в разных областях, в том числе и при анализе информационных потоков [7]. Действительно, теория фракталов и хаоса, которая активно развивается в последнее время, способна на некотором уровне описать структуру информационного пространства.

В настоящее время информационное пространство в целом, ввиду его объемов и динамики изменения, принято рассматривать как стохастическое. Во многих моделях информационного пространства изучаются структурные связи между тематическими множествами, входящими в это пространство. При этом численные характеристики этих множеств подчиняются гиперболическому закону (с возможными степенными поправками). Сегодня в моделировании информационного пространства все чаще используется фрактальный подход, базирующийся на свойстве самоподобия информационного пространства, т. е. сохранение внутренней структуры множеств при изменениях их размеров или масштабов их рассмотрения извне.

Самоподобие информационного пространства выражается, прежде всего, в том, что при почти обвальном росте этого пространства в последние десятилетия, гиперболические частотные и ранговые распределения, получаемые в таких разрезах, как источники и авторы, практически не меняют своей формы. Т. е. применение теории фракталов при анализе информационного пространства позволяет с общей позиции взглянуть на эмпирические законы, составляющие теоретические основы информатики. Например, тематические информационные массивы сегодня представляют развивающиеся самоподобные структуры, и, следовательно, могут рассматриваться как стохастические фракталы. В информационном пространстве возникают, растут и формируются кластеры документов, отражающие современные процессы коммуникации.

Один из основных аспектов самоподобия информационного пространства может быть описан законами Ципфа, выражающими наиболее общие закономерности произвольного текста.

Пусть мы имеем некоторый документ, в котором различные слова присутствуют с различной частотой f. Разделим их на группы так, чтобы в каждой из них находились слова с одинаковой частотой и затем расположим их в порядке убывания частот. Номер группы в этом списке, к которой принадлежит данное слово, называется его рангом r.

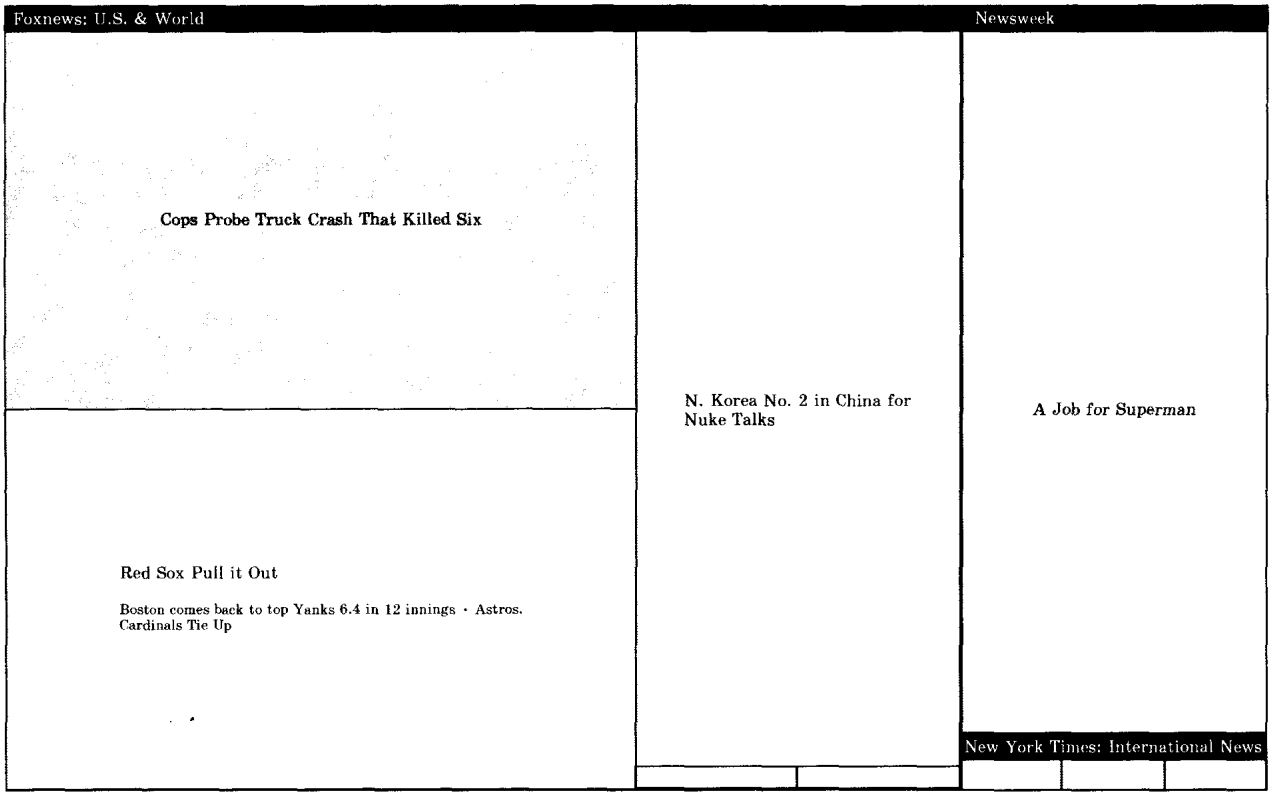
Тогда, согласно первому закону Ципфа, $f\ r = c$, где c — константа, зависящая только от языка, на котором составлен документ (для английского языка, например, $c = 0,1$, а для русского языка $c = 0,07$).

Основатель теории фракталов Бенуа Мандельброт предложил теоретическое обоснование закона Ципфа, полагая, что можно сравнивать язык текста с кодированием. Исходя из требований минимальной стоимости сообщений, Мандельброт математическим путем пришел к аналогичной первому закону Ципфа зависимости $f\ r^e = c$, где e — близкая к единице переменная величина, которая может изменяться в зависимости от свойств текста и языка. Постоянство коэффициента e сохраняется только в центральной зоне диаграммы распределения. По относительной величине той или иной зоны на графике можно судить о характеристиках рассматриваемой в тексте области знаний. Существуют также закономерности, открытые другими учеными (прежде всего, Брэдфордом и Лоткой), являющиеся уточняющими следствиями закономерностей Ципфа и также свидетельствующими о самоподобии информационного пространства.

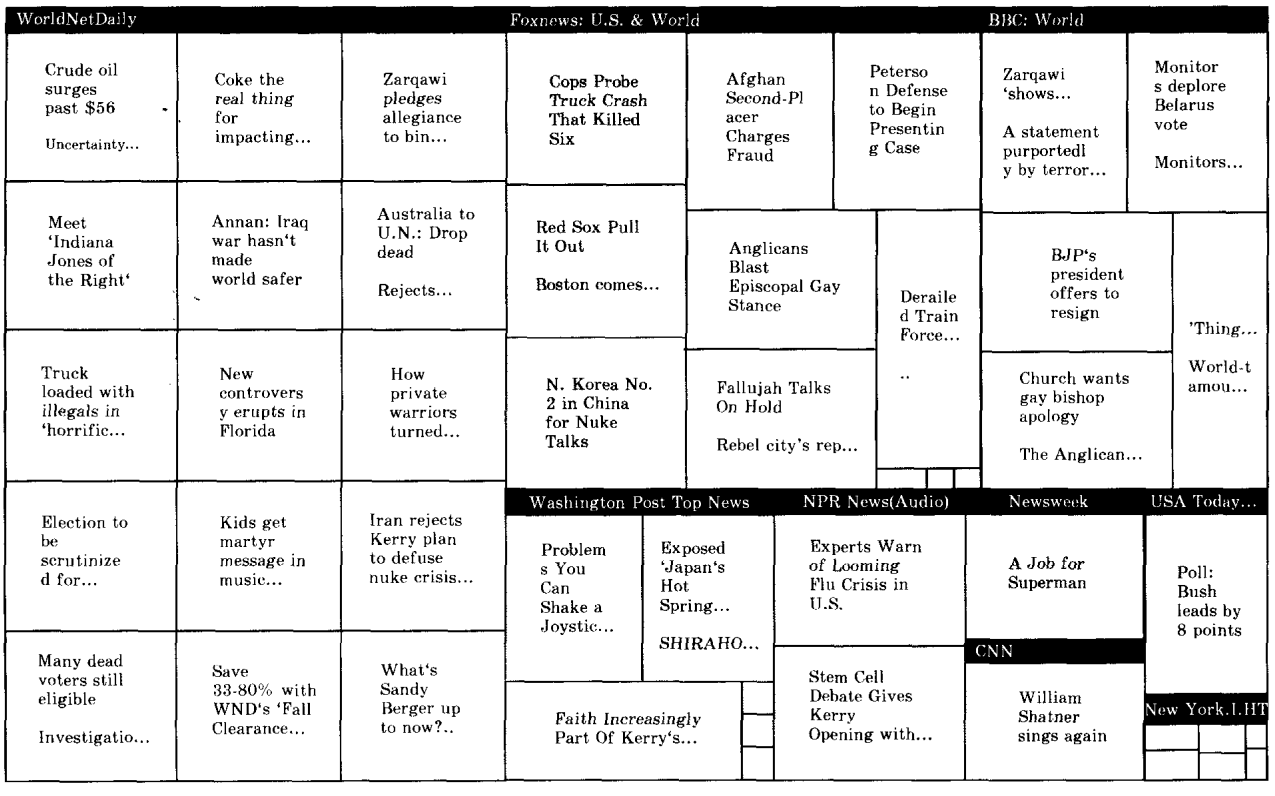
Свойства самоподобия фрагментов информационного пространства наглядно демонстрирует новый интерфейс представленный на веб-сайте службы News Is Free (http://newsisfree.com). На этом сайте отображается состояние информационного пространства в виде ссылок на источники и отдельные сообщения. При этом учитываются два основных параметра отображения — ранг популярности и "свежесть" информации. Укрупненное представление отдельных источников и/или документов — наиболее популярных и актуальных представлено на рис. 4, а.

Средних по популярности документов, конечно, значительно больше. При сохранении общей структуры, происходит "дробление" источников (рис. 4, б).

И наконец, когда предельный ранг популярности, а также "свежести" повышается, дробление уже не позволяет без особых усилий читать названия источников и идентифицировать отдельные документы.



(a)



(б)

Рис. 4. Состояние информационного пространства в соответствии с данными службы News Is Free

При этом, очевидно, последние иллюстрации демонстрируют свойство подобия исследуемой части информационного пространства. Теория фракталов тесно связана с кластерным анализом. Самоподобие информационных кластеров можно рассматривать как сохранение рангового и частотного распределения публикаций по источникам в произвольные моменты, и, кроме того, постоянство параметров этих распределений можно рассматривать как следствие общих структурных закономерностей информационного пространства.

Приведем одну из фрактальных моделей информационного пространства, основанную на подходе, называемом диффузионно-ограниченной агрегацией [8]. Эта стохастическая модель широко применяется для процессов, распространенных в живой природе. Ее обычно определяют следующим образом. Представим себе многомерную сферу (окруж-

ность в двумерном случае) достаточно большого радиуса, на поверхности которой время от времени в случайных местах появляются частицы, которые затем диффундируют внутрь сферы. В центре сферы находится так называемый "зародыш". При столкновении с ним диффундирующая частица "прилипает" к нему и больше не движется (попадает в "архив"). Затем с этим образованием сталкивается следующая, выпущенная с поверхности сферы, частица и так до бесконечности.

В природе так растут кораллы, опухоли, кристаллы. Перенос этой модели на информационное пространство можно интерпретировать следующим образом. Каждой размерности исходной сферы можно приписать определенную тематику, а роль "зародыша" играет исходный информационный массив. При пополнении информационного массива новый документ, размещенный в определенном месте на поверхности сферы стремится к ядру, пересекается с некоторой ветвью и увеличивает ее. Что может дать подобная модель? Самое главное, она может служить эффективным алгоритмом группировки объектов, способным выявлять новые темы (ветви-кластеры), служащие в дальнейшем основой для новой уточненной классификации.

Топология и характеристики модели веб-пространства ("Галстук бабочка") оказались примерно одинаковыми для различных подмножеств веб-пространства, подтверждая тем самым наблюдение о том, что "веб – это фрактал", т. е. свойства структуры всего веб-пространства и его отдельных подмножеств также верны. Таким образом, алгоритмы, использующие информацию о структуре веб-пространства, предположительно должны работать и на отдельных его подмножествах. Информация о структуре веб-пространства уже достаточно широко используется при решении многих задач, например, для оптимизации эффективности механизмов сканирования, при построении новых веб-сервисов, для решения задач анализа и прогноза.

"СКРЫТЫЙ" ВЕБ

Общие принципы организации сетевых структур допускают возможность существования замкнутых областей информационного пространства, недоступных для применения стандартных средств обработки данных. При этом мы не затрагиваем вопрос о том, формируются ли они целенаправленно или возникают случайно. Важно то, что с их возможным наличием необходимо считаться.

Например, данный фактор ставит под сомнение реалистичность оценок темпов роста объема информации в Сети (они могут оказаться намного выше). Нельзя даже исключить возможность того, что видимая часть сетевых данных составляет лишь малую часть от полного их количества.

Не менее серьезной следует считать и проблему возможных потерь данных вследствие попадания их в недоступную область. В перспективе можем получить своего рода "черную дыру", способную неограниченно разрастаться.

Особенно неприятной эта ситуация становится при обработке информационных потоков, поскольку динамично меняющиеся наборы данных значительно сложнее контролировать на различных стадиях их эволюции.

В последнее время появился специальный термин — "Скрытый" веб. "Скрытый" (deep) веб представляет собой часть веб-пространства, недоступную с помощью традиционных информационно-поисковых систем [9]. Большая часть содержания веб-сайтов остается недоступной для поисковых машин, в том числе и потому, что многие веб-серверы хранят и перерабатывают информацию не в том виде, в каком она представляется посетителю. При этом многие веб-страницы генерируются только тогда, когда пользователи обращаются к ним. Традиционные сетевые агенты не умеют работать с подобными ресурсами и не в состоянии определить их содержание. "Скрытый" веб охватывает в первую очередь содержимое онлайновых баз данных. Скрытой является и быстро обновляемая информация — новости, конференции, онлайновые журналы.

В 2000 г. американская компания BrightPlanet (www.brightplanet.com) опубликовала сенсационный доклад, в котором утверждается, что в WWW в сотни раз больше страниц, чем их удалось проиндексировать самыми популярными поисковыми системами.

Нельзя сказать, что не предпринимаются шаги к решению проблем, связанных с замкнутыми областями сети. Существует ряд технологических решений, по крайне мере для определенного класса возникающих задач.

На сегодня разработан целый класс программ, получивших название упаковщиков (wrappers). В некоторых программах, чтобы получить доступ к скрытому содержанию веб-страниц, используется привычный синтаксис поисковых запросов и стандартный формат он-лайн ресурсов. В других системах реализуются преимущества программируемого интерфейса, который позволяет использовать стандартный набор команд и операций.

Для поиска в "скрытой" Сети, а именно в том ее сегменте, который составляют базы данных, сегодня уже существуют некоторые специализированные ресурсы. Среди них, например, системы BigHub (www.bighub.com) и Invisible Web (www.invisible-web.net) компании IntelliSeek. Сайт Invisible Web включает в себя каталог баз данных, большинство из которых не заиндексированы известными поисковыми машинами. При введении запроса этот сайт выдает ссылки на ресурсы, с помощью которых поиск необходимой информации станет наиболее оптимальным. На этом сайте Криса Шермана (Chris Sherman) и Гари Прайса (Gary Price) собраны коллекции ссылок на различные базы данных, среди которых содержится немало уникальных ресурсов, например, сборник спичей политиков и бизнесменов. Программный пакет BullsEye компании IntelliSeek осуществляет поиск более чем в 800 сетевых ресурсах.

В 2005 г. компания Yahoo также запустила тестовую версию поискового сервиса, ориентированного на работу с базами данных сайтов. Он будет проводить поиск не только в общедоступных сайтах, но и на ресурсах, предоставляющих платную информацию, — таких, как онлайновая версия Wall Street Journal, взимающий с посетителей определенную плату. Новый сервис получил название DeepWeb и доступен пока что только для жителей США и Великобритании.

Но все же лидером среди навигаторов в "скрытом" вебе является сайт CompletePlanet (www.completeplanet.com) компании BrightPlanet. Этот сайт — крупнейший каталог, насчитывающий свыше 100 тыс. ссылок. Компания BrightPlanet также создала персональную утилиту для поиска в онлайновых базах данных — LexiBot, которая может обеспечивать поиск в нескольких тысячах поисковых систем "скрытого" веба. Метапоисковый пакет DeepQueryManager (DQM) этой же компании обеспечивает поиск по 55 тысячам "скрытых" Web-ресурсов.

Тем не менее, пока не видно путей решения проблемы в общем виде. Главная сложность состоит в том, что крайне трудно предусмотреть и учесть в явном виде все механизмы утечки данных в замкнутые области, так же, как и все механизмы возникновения и стабилизации таких областей. Поэтому в ближайшее время вряд ли следует ожидать существенных достижений в данном направлении.

ЗАКЛЮЧЕНИЕ

Сегодня необходимо объединение усилий специалистов разных областей. Одна из актуальнейших задач, стоящих перед учеными различных специальностей, состоит в построении четкой модели современного информационного пространства, которая базируется на достижениях в области лингвистики и информатики, а также на методах, близких к методам теоретической физики, на строгом математическом инструментарии. В частности, предполагается дальнейшее развитие обучаемых алгоритмов, которые в противоположность традиционным концепциям искусственного интеллекта должны обеспечить возможность построения рекуррентных процедур с интерактивным участием человеческого интеллекта.

Вместе с тем, исследования современных информационных потоков представляют немалый интерес как для лингвистов, математиков, так и для физиков, например, в плане аналогового моделирования статистических процессов, в том числе сложных нелинейных систем с элементами самоорганизации. Семантика информационного пространства обусловливает и развитие новых методов кодирования и сжатия информации, включая средства обеспечения однозначности дешифровки сообщений.

Предполагается, что новая ступень развития веб-пространства будет определяться технологиями работы с огромным объемом информации, накопившимся в Интернете. Веб следующего поколения будет характеризоваться переходом от сети документов к сети данных, которые при необходимости агрегируются в семантически связанные документы с помощью веб-сервисов. Предполагается также существование единого информационного пространства в виде множества единиц данных, которые могут размещаться на многочисленных сайтах в Интернете. Пользователь будет получать документ путем агрегирования у себя на рабочем месте этих информационных единиц.

Перспективы охвата информационного пространства, по-видимому, будут зависеть от создания и развития эффективной инфраструктуры, в рамках которой будут работать программные продукты со стороны веб-серверов и пользователей.

Даже частичное решение названных задач при наличии обширной и дешевой экспериментальной базы позволит уже в настоящее время реализовать полезные и эффективные инструменты работы и серфинга в информационных потоках.

СПИСОК ЛИТЕРАТУРЫ

1. Ландэ Д. В. Литвин А. Б. Феномены современных информационных потоков // Сети и бизнес.— 2001.— № 1.— С. 14–21.

2. Berners-Lee T., Hendler J., Lassila O. The Semantic Web // Scientific American.— May 2001.— Режим доступа: http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21

3. Григорьев А. Н., Ландэ Д. В. Адаптивный интерфейс уточнения запросов к системе контент-мониторинга InfoStream // Труды Международного семинара "Диалог'2005".— 2005.— С. 109–111.

4. Del Corso G. M., Gulli A., Romani F. Ranking a Stream of News // Proceessings of the 14th International World Wide Web Conference.— 2005.— Режим доступа: www2005.org/cdrom/docs/p97.pdf

5. Ландэ Д. В., Морозов А. Ю. Читайте новости, батенька! // ЧИП-Украина.— 2004.— № 7.— С. 82–85.

6. Broder A. et al. Graph structure in the web / A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, J. Wiener.— Режим доступа: http://www.almaden.ibm.com/cs/k53/www9.final/

7. Иванов С. А. Стохастические фракталы в информатике // НТИ. Сер. 2.— 2002.— № 8.— С. 7–18.

8. Ландэ Д. В. Поиск знаний в Internet.— М.: Диалектика, 2005.— 272 с.

9. Ландэ Д. В. Затерянный вэб // Телеком.— 2005.— № 1.— С. 46–51.

Материал поступил в редакцию 02.09.05.