УДК 681.3

А. Г. Додонов, Д. В. Ландэ, В. Г. Путятин
Институт проблем регистрации информации НАН Украины
ул. Н. Шпака, 2, 03113 Киев, Украина

Современные поисковые технологии — проблемы и некоторые пути их решения

Рассмотрены проблемные вопросы современных поисковых технологий и возможные пути их решения, новые идеи в области сетевого информационного поиска.

Ключевые слова: *информация, аналитическая обработка, Интернет, поиск, ресурс, фильтрация, сбор информации.*

В настоящее время структура, объемы и динамика информационного пространства (прежде всего, Интернет-пространства) обуславливают актуальность поисковых технологий. Большинство пользователей Интернет осуществляет поиск информации с помощью сетевых информационно-поисковых систем (ИПС). Доступ пользователей к современным информационным сетям, эффективное удовлетворение их информационных потребностей возможно только с помощью развитых средств навигации в этих сетях.

Основополагающими характеристиками ИПС являются полнота и релевантность результатов поиска [1]. Полнота поиска тесно связана с оперативностью охвата информации системой. Созданная однажды база данных Интернет-ресурсов является «слепком» состояния Сети в конкретный момент. Если эта база не будет обновляться постоянно и оперативно, то многие из присутствующих в ней ссылок окажутся «мертвыми». Кроме того, отсутствие оперативности обновления баз данных не позволит пользователю отслеживать последние изменения в его предметной области.

Для пользователей ИПС большое значение имеют также такие характеристики как скорость обработки запросов, достоверность отклика (например, оцениваемая по источникам), а также дополнительные сервисы — возможность нахождения документов, подобных уже имеющимся, возможность подключения средств автоматического реферирования и перевода и, конечно же, возможность уточнения запроса.

Поисковые машины следующих поколений должны будут лучше классифицировать информацию и нагляднее представлять ее. Поиск не должен ограничиваться лишь обработкой введенных ключевых слов. Кроме того, имеет смысл пе-

рехода к концепции смысловой навигации в информационных потоках [2] 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 настоящее время основными проблемами в области информационного поиска являются: необходимость охвата больших объемов информации; большая и сложная динамика информационных потоков; многократное дублирование информации; избыток шумовой информации, спама; наличие скрытого веб-пространства, недоступного современным ИПС; отсутствие реальной модели веб-пространства, эффективных алгоритмов поиска в распределенных (например, пиринговых, социальных) сетях, средств смыслового поиска; поиска мультимедийной информации, мультиязычных средств поиска; отсутствие в свободном доступе универсальных поисковых служб, обеспечивающих поиск фактографии, текстовой информации и связей объектов поиска; слабый учет персональных информационных потребностей пользователей; слабая адаптация под эти потребности; явный конфликт при доступе к свободно доступной и/или коммерческой информации. Раскроем некоторые из названных пунктов более подробно.

Необходимость охвата больших объемов информации

В начале существования World-Wide Web небольшое количество веб-сайтов публиковало информацию отдельных авторов для относительно большого количества посетителей. Сегодня с появлением и развитием идеологии Web 2.0 ситуация изменилась. Сами посетители веб-сайтов активно участвуют в создании контента, что привело к резкому росту объема и динамики информационного пространства [3].

Информации в Сети появляется больше, чем ее успевают охватить поисковые системы. Естественно, это влияет на полноту поиска, что объясняет жесткую конкурентную борьбу за объемы проиндексированных веб-документов, ведущуюся поисковыми службами. С самого начала поисковые системы вели ожесточенную борьбу именно за этот показатель. На первых страницах таких поисковых сайтов как Altavista, Google, Alltheweb, Yahoo! публиковались соответствующие цифры — количество проиндексированных документов (объем индекса). В начале XXI века лидером по охвату ресурсов оказалась служба Google. Однако в 2002 году система Alltheweb неожиданно вышла на первую позицию и была признана лучшей сетевой ИПС в мире по охвату ресурсов, проиндексировав 2,1 млрд. веб-страниц. Затем лидерство вновь вернулось Google — свыше 3,3 млрд. веб-страниц в 2003 г. Последняя цифра, размещенная на титульной странице Google в 2005 г., составляла чуть более 8 млрд. страниц. После этого цифры перестали публиковаться. Из официальных пресс-релизов 2005 г. известно, что объем индекса Google составлял 13 млрд. документов, объем индекса Yahoo! превысил это значение и достиг на то время 20 млрд. документов. Администрация Google была не согласна с этой цифрой, выступая с опровержением. Вместе с тем в заявлении Yahoo! было сказано: «Мы поздравляем Google с изъятием с их главной страницы

числа, показывающего размер индекса, и с признанием того, что оно ничего не значит. Как мы уже говорили, важно лишь, чтобы потребители находили то, что они ищут, и мы предлагаем пользователям сравнить результаты поиска наших систем».

Казалось бы, возвращаться к оценке объема индекса никто не будет. Однако в июле 2008 года появилась новая глобальная поисковая система Cuil с относительно небольшим бюджетом (33 млн. долларов), содержащая в индексе 121 млрд. веб-страниц, что, по мнению экспертов, в несколько раз превышало индекс Google, который официально не обнародовался. Можно лишь косвенно сравнивать показатели Google и Cuil, задавая им простейшие запросы (информации Cuil можно доверять — ее создатели предъявили поисковый индекс внешним экспертам). Как явствует из материалов компаний, обе поисковые системы не используют так называемого стоп-словаря, т.е. запросы по простым, часто употребляемым словам позволят оценить соотношение объемов индексов. И такую оценку с определенным уровнем достоверности может сделать каждый [4]. Например, введя поисковое слово «the» одновременно двум системам, можно получить:

Google: about **22,550,000,000** for **the**;

Cuil: 22,883,636,124 results for **the**.

Результаты вполне сопоставимы — можно сделать вывод о примерно одинаковом объеме поисковых индексов. Введем слово «для» (для проверки русскоязычной части), получаем:

Google: about **546,000,000** for **для**;

Cuil: 368,508,113 results for **для**.

Русскоязычная часть индекса Google оказалась несколько большей. О низком качестве (объеме) русскоязычного индекса Cuil свидетельствуют и запросы по другим словам.

Неожиданный результат получается для еще одного слова — «of»:

Google: about **22,760,000,000** for **of**;

Cuil: 121,000,000,000 results for **of**.

В этом случае у Cuil результат более чем в 5 раз весомей. Но, учитывая итоги поиска по слову «the» (и по другим словам, в частности, не только на английском языке), можно сделать иной вывод. Каковы бы ни были результаты подобных сравнений, факт остается фактом: Google — самая популярная поисковая система, самый дорогой бренд в мире, а Cuil — мало кому известный проект с бюджетом региональной поисковой системы. Это подтверждает тот факт, что ситуация на рынке поисковых систем не простая — она отражает принцип новой экономики: здесь не может быть вторых ролей. Или система лучшая в мире, или ней никто не будет пользоваться. Система должна найти свою нишу в задаче максимального удовлетворения запросов пользователей — быть самой полной, самой демократичной, самой интеллектуальной или самой локализированной.

Дублирование информации

Документы, публикуемые на веб-сайтах, зачастую многократно дублируются в виде перепечаток или пересказов. Практически все сетевые ИПС содержат компоненты определения содержательного дублирования. Однако достижение при-

емлемого качества выявления подобных документов (дубликатов) при использовании различных критериев является открытой научно-прикладной проблемой. Задача выявления дубликатов, а также перепечаток документов с небольшими изменениями («почти дублей») является одной из актуальнейших и сложнейших при интеграции информационных ресурсов. Существующие в настоящее время алгоритмы выявления дублей в современных информационных потоках требуют применения самых современных компьютерных комплексов, содержащих тысячи серверов (что можно видеть на площадках современных сетевых поисковых служб), суперкомпьютеров.

Если нахождение явно дублирующейся информации не представляет проблем, то смысловые дубликаты выявляются не так легко, здесь на помощь приходят алгоритмы сопоставления и вероятностной оценки содержимого документов. Кроме того, Интернет является «агрегатором» информации, не находящейся в открытом доступе.

Было проведено исследование того, в какой мере платные информационные материалы, доступные платным подписчикам основных информационных агентств (ИА) Украины и России, становятся доступными в открытом доступе на информационных веб-сайтах [5]. Ведь в этом случае ценность информационных сообщений во многом определяется оперативностью, поэтому отдельной задачей была оценка запаздывания публикаций в Интернет по сравнению с временем рассылки соответствующих сообщений.

При проведении исследований авторы получили уникальную возможность доступа к подписным материалам ведущих ИА, представленных в украинском информационном пространстве. Кроме того, в распоряжении авторов находилась система контент-мониторинга InfoStream [6] — поисковая система, с помощью которой в реальном масштабе времени сканируется свыше 3000 информационных веб-сайтов, представленных в украинском и российском сегментах веб-пространства. Таким образом, в ходе исследования рассматривались два текстовых корпуса — сообщений ИА и текстов, сканированных из веб-пространства. Рассматривались сообщения ИА по общеполитической тематике, поступающие в течение 20 дней одного месяца. Эти сообщения сравнивались с текстами, сканируемыми из Интернета в течение всего месяца, количество которых составило свыше 1 млн. документов.

Технически задача нахождения дубликатов (в данном случае речь идет именно о дубликатах, а не о сообщениях по той же теме, перепечаток с незначительными искажениями) решалась методом, описанным в [7]. Этот метод относится к группе методов нахождения «подобных» документов [8–10], основанных на выделении некоторого множества опорных слов, имеющих наибольший TFIDF [11]. В качестве некоторых «инвариантов» для сообщений использовались цепочки из 12 опорных слов, прошедших процедуру морфологической обработки. Такое небольшое количество опорных слов объясняется небольшой средней длиной новостных сообщений (2000–3000 символов).

В результате проведенных исследований удалось получить такие данные:

— 62 % сообщений ИА было опубликовано на веб-сайтах;

— общее количество перепечаток на различных веб-сайтах составило 456 %!;

— количество перепечаток с положительным временем запаздывания (из материалов ИА — на веб-сайты) составило 73 %;

— количество перепечаток с отрицательным временем запаздывания (перепечаток из Интернет, помещаемых в ленты ИА) составило 27 %.

Ранжированный график распределения сообщений ИА по времени задержки публикаций приведен на рис. 1, на котором четко видны экстремальные отклонения в начальной и конечной области.

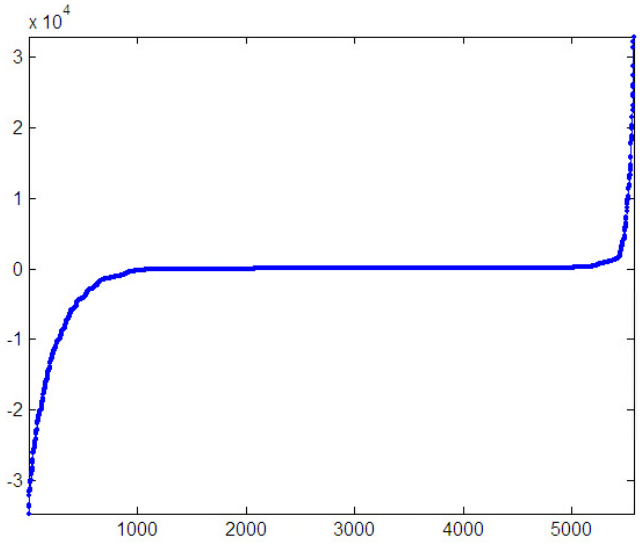

Рис. 1. Распределение сообщений ИА (ось абсцисс) по времени запаздывания в минутах (ось ординат)

Отклонение в начальной области характеризует большое время задержки включения в ленты ИА материалов, размещенных, как правило, на сайтах органов государственной власти, что объясняется инертностью ИА, отсутствием у них средств мониторинга веб-пространства. Отклонения в конечной области объясняются задержками перепечаток на веб-сайтах сообщений, получивших со временем некоторое новое продолжение. Вместе с тем центральная область графика имеет стабильный характер со средним значением около получаса.

Массовый характер перепечаток позволяет делать выводы о том, что все сообщения, интересные пользователям веб-сайтов, были перепечатаны. По-видимому, примерно 37 % сообщений ИА оказались им недостаточно интересными.

Результаты исследований заставили задуматься, за что же платят подписчики информационным агентствам сегодня, когда большая часть информации с минимальной задержкой доступна в Интернет? По-видимому, за аналитический подбор этой информации, репрезентативность и достоверность.

Динамика информационных потоков

Новый уровень развития сетевого информационного пространства обуславливает необходимость создания и развития адекватных моделей информационных потоков. В этой связи возникает интерес к подходам, основанным на понимании информации как меры упорядоченности некоторой системы и, соответственно, к статистическим методам ее обработки. Для организации эффективной коммуни-

кации в сетях сегодня приходится постоянно возвращаться к истокам теории информации, понятиям энтропии, теории Шеннона, уравнениям Больцмана и др., что обуславливает широкие перспективы применению мощного аппарата математики и физики в решении теоретико-информационных задач [2].

При моделировании этих процессов используются методы нелинейной динамики, теории клеточных автоматов и самоорганизованной критичности. При моделировании информационных потоков изучаются структурные связи между входящими в них массивами документов. Сегодня при этом все чаще применяется фрактальный анализ, подход, базирующийся на свойствах сохранения внутренней структуры массивов документов при изменениях их размеров или масштабов рассмотрения. Теория информации, которая ранее находила свое основное применение в области передачи данных, становится полезной и для анализа текстовых массивов, динамически порождаемых в сетях.

Введем формальное определение информационного потока [1], которое корреспондируется с классическим определением из теории информации. Рассмотрим отрезок (a,τ) оси времени, где $\tau > a$. Допустим, за этот интервал времени в соответствии с некоторыми закономерностями в сети публикуется некоторое количество сообщений (документов) — k. На оси времени моменты публикации отдельных сообщений обозначим как $\tau_1,\ \tau_2\ ,...,\ \tau_k\ \ (a \le \tau_1 \le \tau_2 \le ... \le \tau_k \le \tau)$. Информационным потоком будем называть процесс $N_\alpha(\tau)$, реализация которого характеризует количество сообщений, появившихся в интервале (a,τ), как функцию правого конца отрезка τ. В соответствии с этим определением реализация информационного потока является неубывающей ступенчатой всегда целочисленной функцией $N_\alpha(\tau)$.

Приведенное определение на локальных временных областях соответствует действительности, но не учитывает эффект старения информации, противоречащий «накопительной» способности информационного потока $N_\alpha(\tau)$ на больших промежутках времени. Определенный выше информационный поток учитывает лишь количество информационных сообщений, вне зависимости от их содержания, определение которого является достаточно субъективным процессом. Для строгого моделирования тематических информационных потоков используют модели, которые отличают документы по отдельным словам или словосочетаниям (обычно их называют термами, от англ. *Terms*).

В традиционной сетевой ИПС информационное пространство, которое состоит из стабильной и динамической частей, и индексируется с помощью ИПС, изменяет свое наполнение во времени: некоторые новостные документы поступают в стабильную часть в виде архивов, а другие исчезают. В этом случае пользователь при обращении к ИПС находит релевантные запросу документы из стабильной части, ссылки из динамической части, которые устарели, и ничего не находит из обновленной динамической части.

В настоящее время ни одна из традиционных ИПС в достаточном объеме не помогает в поиске актуальной новостной информации, которая находится в динамической части сети Интернет. Решение этой задачи требует применения посредника — системы интеграции информационных потоков. Принцип индексирова-

ния, которое должно осуществляться этим посредником, немного отличается от индексирования традиционными поисковыми системами: должен индексироваться не весь контент сети, а только динамическая часть. В результате такого подхода пользователь будет получать необходимые ответы из новостной и из «устаревшей» новостной части (подтвержденных документами из архивной БД), но не получит полной выборки документов из стабильной части Интернет. Таким образом, проблема достижения полноты при поиске в динамической сети может быть решена путем использования двух инструментов — традиционных ИПС (для стабильной части веб-пространства) и систем интеграции информационных потоков.

В настоящее время задачи мониторинга информационных потоков в компьютерных сетях, их адаптивного агрегирования и обобщения осложняются отсутствием типовых методик и решений, неполнотой существующих технологических подходов. Вместе с тем, опыт создания и внедрения корпоративных информационных систем свидетельствует о необходимости создания и внедрения документальных информационных хранилищ для обеспечения научных исследований, получения разнообразных аналитических сведений, навигации в документальных информационных потоках больших объемов [12, 13].

Избыток шумовой информации, спама

Шумовая информация, зачастую несанкционированно навязываемая пользователям (не только электронной почты, но и других сетевых сервисов), в последнее время получила название «спам». Проблема спама породила две задачи — задачу его выявления и задачу извлечения небольшого количества информации, действительно необходимой пользователю [14].

Сегодня спам выявляется с помощью сложных комбинированных алгоритмов, требующих, как правило, мощных вычислительных ресурсов. Вместе с тем, смысловой основой этих алгоритмов является так называемый «наивный» метод Байеса, в рамках которого подразумевается использование оценочной базы — двух текстовых корпусов (например, электронных писем), один из которых составлен из спама, а другой — из полезных документов. Для каждого из корпусов подсчитывается частота использования каждого слова, после чего вычисляется весовая оценка (от 0 до 1), характеризующая условную вероятность того, что сообщение с этим словом является спамом. Значения весов, близкие к ½, не учитываются при интегрированном расчете, поэтому слова с такими весами игнорируются и удаляются из словарей.

В соответствии с методом, предложенным Полом Грэмом, если сообщение содержит n слов с весовыми оценками w_1, ..., w_n, то условная вероятность того, что письмо окажется спамом, основанная на данных из оценочных корпусов, вычисляется по формуле:

$$Spm = \Pi w_i / (\Pi w_i + \Pi(1 - w_i)).$$

Эта формула обосновывается следующим соображением. Предполагается, что S — событие, заключающееся в том, что письмо — спам; A — событие, заключающееся в том, что письмо содержит слово t. Тогда, в соответствии с форму-

лой Байеса, справедливо

$$P(S \mid A) = \frac{P(A \mid S)P(S)}{P(A \mid S)P(S) + P(A \mid \overline{S})P(\overline{S})}.$$

Если изначально не известно, является письмо спамом или нет, исходя из опыта, предполагается, что $P(\overline{S}) = \lambda P(S)$, на основании чего следует:

$$P(S \mid A) = \frac{P(A \mid S)}{P(A \mid S) + \lambda P(A \mid \overline{S})}.$$

Далее полученная формула обобщается следующим образом. Предполагается, что A_1 и A_2 — это события, заключающиеся в том, что письмо содержит слова t_1 и t_2. При этом вводится допущение, что эти события независимы (именно поэтому метод называется «наивным» байесовским). Условная вероятность того, что письмо, содержащее оба слова (t_1 и t_2) является спамом, равна:

$$P(S \mid A_1 \,\&\, A_2) = \frac{P(A_1 \mid S)P(A_2 \mid S)}{P(A_1 \mid S)P(A_2 \mid S) + \lambda P(A_1 \mid \overline{S})P(A_2 \mid \overline{S})} =$$

$$= \frac{p(t_1)p(t_2)}{p(t_1)p(t_2) + \lambda(1 - p(t_1))(1 - p(t_2))}.$$

Обобщением этой формулы на случай произвольного количества слов и $\lambda = 1$ является формула П. Грэма. Широкое применение в антиспамовских фильтрах находит именно значение $\lambda = 1$. Это допущение упрощает вычисления, но искажает действительность и существенно снижает качество работы соответствующих программ.

Скрытое веб-пространство

По оценкам экспертов, с помощью даже самых крупных глобальных поисковых систем в Интернете сегодня доступно не более 30 % открытой информации, присутствующей в веб-среде. Веб-ресурсы, находящиеся в свободном доступе, но не доступные с помощью обычных поисковых систем, образуют так называемый глубинный веб. Эти ресурсы имеют собственное название — «глубинный» или «скрытый» (deep) веб, которое ввел Джилл Иллсворт (Jill Ellsworth) в 1994 году, обозначив им документы, недоступные для обычных поисковых систем [3]. Интегрированный доступ к таким ресурсам все еще остается открытой проблемой, частичное решение которой достигается с помощью специальных каталогов и систем, зачастую доступных обычным пользователям Интернета. Глубинный веб чаще всего охватывает динамически формируемые веб-страницы, содержание которых хранится в базах данных и доступно лишь по запросам пользователей. Иногда для доступа к подобным страницам используется так называемый тест Тьюринга (или тест на разумность): предлагается решить арифметическую задачу,

загадку или попросту ввести в определенное поле последовательность символов, изображенную графически. К примеру, БД с законодательными документами Украины или России (системы «Рада», «Кодекс», соответственно) вполне можно отнести к такой категории, ведь размещенные в них сотни тысяч документов, доступные для свободного просмотра, не попадают в индексы глобальных сетевых ИПС.

Основатель BrightPlanet Майкл Бергман выделил 12 разновидностей «скрытых» веб-ресурсов, в списке которых оказались как традиционные базы данных (патенты, медицина и финансы), так и публичные ресурсы — объявления о поиске работы, чаты, библиотеки, справочники. Бергман причислил к «скрытым» ресурсам и специализированные поисковые системы, обслуживающие определенные отрасли или рынки, базы данных которых не включаются в глобальные каталоги традиционных поисковых служб. К «скрытому» вебу также относятся многочисленные системы интерактивного взаимодействия с пользователями — помощи, консультирования, обучения, требующие участия людей для формирования динамических ответов от серверов. К ним также можно отнести и закрытую (полностью или частично) информацию, доступную пользователям только с определенных адресов или групп адресов, иногда городов или стран. К «скрытой» части веб-пространства многие причисляют и веб-страницы, зарегистрированные на бесплатных серверах, которые индексируются, в лучшем случае, лишь частично. В поисковых системах обычно ограничивается глубина индексирования таких сайтов во избежание возможного циклического просмотра одних и тех же страниц.

Поиск в P2P (пиринговых) сетях

В настоящее время веб-пространство не является крупнейшим информационным ресурсом в Интернете. Основной объем ресурсов сосредоточен в «пиринговых» сетях (P2P — «точка – точка»), многие из которых являются так называемыми «файлообменными» [15]. В таких сетях отсутствуют выделенные серверы, а каждый узел является как клиентом, так и сервером. Пиринговые сети состоят из узлов, каждый из которых взаимодействует лишь с некоторым подмножеством других узлов. При освещении этой тематики учитывалось то, что проблемы поиска и уязвимости в таких сетях до сих пор остаются открытыми.

Существует несколько областей применения пиринговых сетей, объясняющих их растущую популярность, назовем некоторые из них.

Обмен файлами. P2P выступают альтернативой FTP-архивам, которые утрачивают перспективу из-за значительных информационных перегрузок.

Распределенные вычисления. Например, такой проект с элементами P2P как SETI@HOME, посвященный распределенному поиску внеземных цивилизаций, продемонстрировал высокий вычислительный потенциал для распараллеливаемых задач. Вместе с тем, этому проекту свойственна централизованная раздача и прием данных.

Обмен сообщениями. Как известно, ICQ — это P2P-проект. Эта сеть также обладает элементами централизации, в частности, очень зависит от состояния сервера login.icq.com.

Интернет-телефония. Сегодня одной из самых популярных служб Интернет-телефонии является Skype (www.skype.com), созданная в 2003 г. Н. Зеннстромом и Я. Фриисом, авторами известной пиринговой сети KaZaA. Построенная в архитектуре P2P служба Skype охватывает свыше 10 млн. пользователей.

Групповая работа. Сегодня реализованы такие сети групповой работы, как Groove Network (защищенное пространство для коммуникаций) и OpenCola (поиск информации и обмен ссылками).

Вопрос эффективного поиска в таких сетях остается открытым, существуют лишь специальные поисковые сайты в веб-пространстве, помогающие решить эту проблему.

На практике пиринговые сети состоят из рабочих станций, каждая из которых взаимодействует лишь с некоторым подмножеством узлов сети (из-за ограниченности ресурсов). Достаточно часто пиринговые сети дополняются выделенными серверами. Такие серверы позволяют решать вопросы поиска по запросам, так как именно эта проблема для пиринговых сетей не может считаться решенной.

Файлообменные P2P-сети уже в начале 2010 г. охватывали более 150 млн. узлов. Сегодня в Интернет более половины всего трафика приходится на файлообменные P2P-сети. Наиболее популярные из них — это Bittorrent, Gnutella2 и eDonkey2000.

При поиске в пиринговых сетях тема полноты поиска отодвигается на второй план, главная же задача — быстрое и эффективное нахождение наиболее релевантных откликов на запрос, передаваемый от рабочей станции всей сети. В частности, актуальная проблема — уменьшение сетевого трафика, порождаемого запросом (например, пересылки запроса по многочисленным рабочим станциям), и в то же время получение наилучших характеристик выдаваемых документов, т.е. получение качественного результата.

Приемлемое качество поиска в пиринговых сетях на сегодняшний день обеспечивают лишь специализированные, централизованно наполняемые, поисковые веб-сайты, работающие по протоколу HTTP. Например, для файлообменной сети eMule таким поисковым сервером является сайт Figator.com, а для сети Bittorrent — сайт isoHunt.com.

Как и для файлообменных сетей, для этих серверов особо актуальными и критичными являются проблемы качества и достоверности предоставляемого контента, фальсификация файлов и распространение фальшивых ресурсов, вирусов, «троянских коней», возможность фальсификации ID рабочих станций.

Существует несколько алгоритмов поиска в таких сетях [16], ни один из которых не подходит для получения результатов, сравнимых с даже традиционным поиском в веб-пространстве. Наиболее популярные алгоритмы базируются на поиске ресурсов по ключам. В большинстве пиринговых сетей, ориентированных на обмен файлами, используются два вида сущностей, которым приписываются соответствующие идентификаторы (ID): узлы и ресурсы (например, файлы), которые характеризуются ключами (Key), то есть сеть может быть представлена двумерной матрицей размерностью MN, где M — количество узлов, N — количество ресурсов. В этом случае задание поиска сводится к нахождению ID узла, на котором сохраняется ключ ресурса. Одним из наиболее эффективных алгоритмов поиска в сетях P2P является так называемый «Интеллектуальный поисковый меха-

низм» (*Intelligent Search Mechanism, ISM*). Улучшение скорости и эффективности поиска информации с помощью данного метода достигается за счет минимизации расходов на количество сообщений, которые передаются между узлами, а также количества узлов, которые опрашиваются для каждого запроса. То есть оцениваются лишь те узлы, которые больше всего отвечают конкретному запросу.

ISM состоит из двух компонент — профайла и способа его ранжирования, так называемого ранга релевантности. Каждый узел сети строит информационный профайл для каждого из соседних узлов. Профайл содержит последние ответы от каждого из узлов. С помощью ранга релевантности осуществляется ранжирование профайлов узлов для выбора тех соседних, которые будут давать наиболее релевантные документы по запросу.

При реализации модели ISM применяется единый стек запросов, в котором сохраняется по T запросов для N узлов. Как только стек заполняется, происходит замена того запроса, который не использовался дольше (*Least Recently Used*, *LRU*), с целью сохранения последних запросов. Функция «ранг релевантности» (*Relevance Rank, RR*) применяется узлом P_l, чтобы выполнять оперативную классификацию его соседей для определения тех из них, которые стоит опрашивать первыми по запросу q. Для вычисления ранга релевантности каждого узла P_i, P_l сравнивает q со всеми запросами в структуре профайла, для которого известен список ответов на предыдущие запросы, и вычисляет $RR(P_i, q)$:

$$RR(P_i, q) = \sum_{j \in Q} Sim(q_j, q)^{\alpha} \cdot S(P_i, q_j),$$

где α — параметр, который задает вес запросов. В этой формуле Q — множественное число запросов, на которые был ответ от узла P_i; $S(P_i, q_j)$ — количество результатов, которые возвращались узлом P_i по запросу q_j; метрика *Sim* рассчитывается по правилу, принятому в векторно-пространственной модели поиска:

$$Sim(q_j, q) = \frac{q_j \cdot q}{|q_j||q|}.$$

Ранг релевантности *RR* обеспечивает более высокий ранг узлу, который возвращает больше результатов.

Метод ISM эффективно работает в сетях, узлы которых содержат некоторые специализированые сведения. В частности, исследование сети Gnutella показывает, что качество поиска очень зависит от «окружения» узла, с которого задается запрос. Большая проблема в методе ISM заключается в том, что поисковые сообщения могут циклически проходить через те же узлы сети, не достигая некоторых ее частей. Чтобы решить эту проблему для охватывания большей части сети, предложен подход, при котором для каждого запроса выбиралось небольшое подмножество случайных узлов, которые добавлялись к набору релевантных узлов.

Существуют также другие подходы к решению этой проблемы, например, применяемый в протоколе BGP4 (RFC 1771), где каждый запрос хранит «историю» — список узлов, через которые он уже прошел.

Необходимость создания новой модели веб-пространства

Эффективный анализ информационных потоков в Интернет, построение эффективных ИПС невозможно без некоторых сведений о структуре самого веб-пространства. В 1999 г. А. Брёдер из IBM и его соавторы из компаний AltaVista, IBM и Compaq сделали первую попытку математического описания «карты» ресурсов и гиперсвязей веб-пространства, получившей благодаря своей форме название «галстука-бабочки» (Bow Tie). С помощью баз данных и поискового механизма AltaVista было проанализировано свыше 200 млн. веб-страниц и несколько миллиардов ссылок, размещенных на этих страницах [2].

В рамках общей задачи определения структуры связей между отдельными веб-страницами было выявлено: центральное ядро (28 % веб-страниц) — зона сильной связности сети (*Strongly Connected Component, SCC*); «отправные веб-страницы» (IN), охватывающие 22 % ресурсов; «конечные веб-страницы» (OUT), также охватывающие 22 % ресурсов; «отростки, мысы и перешейки» (22 % веб-страниц). Существуют и «острова», которые вообще не пересекаются с остальными ресурсами Интернет.

Было обнаружено, что пропорции названных категорий в течение нескольких месяцев оставались неизменными, несмотря на значительное увеличение общего объема веб-ресурсов. Топология и характеристики модели оказались примерно одинаковыми для различных подмножеств веб-пространства, подтверждая тем самым наблюдение о том, что свойства структуры всего веб-пространства Bow Tie также верны и для его отдельных подмножеств. Таким образом, алгоритмы, использующие информацию о структуре веб-пространства, предположительно будут работать и на отдельных его подмножествах [1].

Оказалось, что распределение степеней узлов (входящих и исходящих гиперссылок) веб-пространства (исследовались сайты домена edu в количестве 325729) подчиняется степенному закону, т.е. вероятность того, что соответствующая степень вершины равна i, пропорциональна $1/i^k$ (для входящих ссылок $k \approx 2{,}1$, а для исходящих $k \approx 2{,}45$). Кроме того, оказалось, что сеть WWW является «малым миром» со средней длиной кратчайшего пути, равной 11, и относительно большим значением коэффициента кластерности, приблизительно равным 0,15 (для классического случайного графа это значение составило бы 0,0002) .

Вместе с тем необходимо подчеркнуть некоторую некорректность расчета объемов «островов» по Брёдеру из-за того, что список веб-ресурсов был получен из БД системы AltaVista, полученный в результате работы программы-робота, сканирующего веб-ресурсы, переходя от одного к другому по гиперссылкам.

Модель Брёдера не учитывает особенностей динамической части веб-пространства, формируемой потоками новостных сообщений [17]. Применение модели «галстука-бабочки» к динамической составляющей веб-пространства нельзя считать корректным по ряду причин:

— динамика информационных потоков влияет на природу гиперссылок, на сообщения, например, в течение определенного времени их может вообще не существовать;

— модель Брёдера слабо учитывает особенности «скрытого» Web;

— в информационных потоках необходимо учитывать не только гиперссылки, но и ссылки контекстные, причем не только на объекты из открытой части веб-пространства;

— модель Брёдера не включает такого понятия как смысловое дублирование информации;

— за прошедшее время с момента создания модели Брёдера появились новые разновидности гиперсвязей в веб-пространстве, например, существуют гиперссылки, доступные для пользователей-людей, но недоступные для роботов поисковых систем (в частности, определяемые тегом <noindex>).

Проблемы смыслового поиска

Для пользователя пертинентность, соотношение объема полезной для него информации к общему объему полученной информации, имеет решающее значение. При этом следует учитывать, что формальный запрос к системе является предметом творческого осмысления информационной потребности и не всегда точно отражает последнюю. Достижение высокой пертинентности — основное поле конкурентной борьбы современных поисковых систем. Именно для максимального удовлетворения информационных потребностей пользователей поисковые системы сегодня максимально интеллектуализируются, получили широкое практическое применение теории и методы семантических сетей, контент-анализа и глубинного анализа текстов (Text Mining).

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

Следует признать, что многие основные задачи Семантического Web в настоящее время выглядят достаточно химерными. Вместе с тем частные решения, полученные при попытках реализации Семантического Web, сегодня широко применяются в информационных технологиях. К таким решениям относятся, например, агрегация новостей или ведение блогов (интерактивных сетевых журналов) на основе XML.

Text Mining

Поиск в сетевой среде может стать более эффективным за счет технологий глубинного анализа текстов (Text Mining), нахождения в текстах аномалий и трендов.

Концепция глубинного анализа текстов Text Mining включила в себя технологические и методологические подходы контент-анализа, компьютерной лингвистики [2], в частности, автоматическое реферирование, анализ взаимосвязей понятий, построение поисковых образов документов.

Разработанные на основе статистического и лингвистического анализа, а также методов искусственного интеллекта, технологии Text Mining предназначены для проведения смыслового анализа. Задача Text Mining — выбирать из текстов наиболее ключевую и значимую информацию для пользователей. Важная компонента технологий Text Mining связана с извлечением из текста характерных элементов или признаков, которые могут использоваться в качестве ключевых слов, метаданных, аннотаций. Еще одна задача Text Mining — отнесение документов к некоторым категориям из заданной схемы их систематизации. Кроме того, Text Mining — это новый вид поиска, который в отличие традиционных подходов не только находит списки документов, формально релевантных запросам, но и помогает в понимании смысла текстов. Таким образом, пользователю будет незачем самому «просеивать» огромное количество неструктурированной информации. Text Mining — это алгоритмическое выявление прежде не известных связей в уже имеющихся данных. Применяя Text Mining, пользователи могут получать новую ценную информацию — знания.

В соответствии с уже сложившейся методологией, к основным элементам Text Mining относятся: классификация, кластеризация, извлечение фактов, понятий, реферирование, ответ на запросы, тематическое индексирование и поиск по ключевым словам.

Извлечение понятий из текста представляет собой технологию, обеспечивающую получение информации в структурированном виде. В качестве структур рассматриваются как относительно простые понятия (ключевые слова, персоны, организации, географические названия), так и более сложные, например, имя персоны, ее должность в конкретной организации и т.п. Данная технология включает три основных подхода:

1) Entity Extraction — извлечение слов или словосочетаний, важных для описания содержания текста. Это могут быть списки терминов предметной области, персон, организаций, географических названий и др.;

2) Feature Association Extraction — прослеживание связей между извлеченными понятиями;

3) Event and Fact Extraction — извлечение сущностей, распознавание фактов и событий.

Следует отметить, что подходы к извлечению различных типов понятий из текстов существенно разняться как по контексту их представления, так и по структурным признакам. Так, для выявления принадлежности документа к тематической рубрике могут использоваться специальным образом составленные запросы на информационно-поисковых языках, включающих логические и контекстные операторы, скобки и т.д. Выявление топонимов предполагает использование таблиц, в которых кроме шаблонов написания этих названий, используются коды и названия стран, регионов и отдельных населенных пунктов.

Необходимость универсальных поисковых служб

Существующие доступные фактографические базы данных структурированной информации не всегда могут прийти на помощь исследователю-аналитику. Для оперативного определения фактов и сущностей, моделирования информационных связей между ними наиболее перспективным подходом оказывается учет информации, знаний, которые содержатся в неструктурированных текстовых документах, в частности, в Интернет. Поиск в массивах неструктурированной текстовой информации может применяться для задач наведения исследователей-аналитиков «на цель» в условиях, когда фактографические базы данных структурированной информации труднодоступны, неполны, неоперативны.

Неструктурированные тексты содержат в себе несравненно больше важной информации, чем структурированные записи баз данных, именно в силу того, что формализации подлежит сравнительно небольшой сегмент информации. В настоящее время появляется все больше качественных инструментальных средств извлечения понятий из неструктурированных текстов.

Сегодня, когда у пользователей уже накоплен большой опыт работы с традиционными ИПС, оказалось очевидным, что факты или понятия, которые ищутся с помощью таких систем, сами по себе зачастую бессмысленны. Например, если пользователя интересуют информационные связи Сбербанка России с другими банками или частными лицами, то он не знает, какие банки или фамилии ему указать в запросе, а охватить все документы, содержащие словосочетание «Сбербанк России» физически невозможно. В таких случаях информационные связи, интенсивность которых выходит за рамки статистического фона, как правило, отражают реальность.

Интерпретируют обычно не сами понятия или факты, а взаимосвязи между ними. «…Важным оказывается не столько исследование самих понятий, сколько исследование их взаимосвязи. Именно взаимосвязь способствует пониманию мотивационно-целевых особенностей отношений человека...» [18]. То есть пользователя интересует не понятие само по себе, а понятие в окружении, что позволяет сразу иметь представление о предметной области, при необходимости направить уточняющий поиск в нужном направлении. Элементы такого подхода можно видеть, например, в «облаках» системы Quintura (http://quintura.ru), но там отображаются не понятия/сущности, а наиболее часто используемые термы.

Таким образом, объективно существует необходимость построения эффективной полнотекстовой ИПС, обеспечивающей поиск не по отдельным термам или понятиям, а по взаимосвязям между сущностями, присутствующими в документах, то есть создания систем, которые будем условно называть «базами данных связей» (БДС).

В корпоративной информационной инфраструктуре база данных связей может использоваться различным образом, например, отдельно, либо возможности БДС могут быть дополнены возможностями существующих полнотекстовых и/или фактографических баз данных (рис. 2). При этом основным результатом работы БДС является построение карт связей, а в качестве побочного эффекта, реализующего «режим доказательства», может рассматриваться извлечение самих документов как источников связей.

При проектировании БДС должны использоваться решения, которые можно отнести к самым перспективным в области создания информационно-аналитических систем, в частности, теория и технологии глубинного анализа текстов — Text Mining [2], в том числе развитая методология экстрагирования понятий [19], теория и технологии баз данных сверхбольших объемов, концепция «сложных сетей» (complex networks) [16]. Теория сложных сетей изучает характеристики, учитывая не только на топологию сетей, но и статистические феномены, распределение весов отдельных вершин (в качестве которых можно рассматривать сущности, понятия, факты) и ребер, эффекты протекания и проводимости в сетях и т.п.

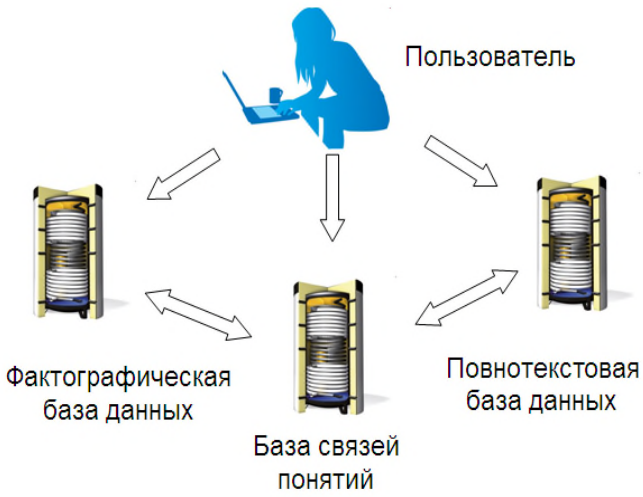

Рис. 2. Место базы данных связей понятий в корпоративной информационной инфраструктуре

Анализируя связи в сети, можно определить многие неочевидные свойства, например, выявить наличие кластеров, определить их состав, различия в связности внутри и между кластерами, идентифицировать ключевые элементы, которые связывают кластеры между собой и т.п. Серьезным препятствием при анализе является неполнота информации о связях между отдельными узлами сети. Вместе с тем сегодня уже существуют алгоритмы, с помощью которых становится возможным с высокой вероятностью восстановить отсутствующие фрагменты связей. Даже не имея полного описания информационной сети, можно получать репрезентативную выборку «реальных» связей и по ней достроить всю сеть.

Учет персональных потребностей пользователей

Представляется очень важным, чтобы агрегирование информации, обеспечение доступа пользователей к этой информации было адаптивным, т.е. ориентированным на информационные потребности конкретных пользователей. Если учитывать динамику и объемы доступной информации в Интернет, то становится очевидным, что обеспечение эффективного доступа в режиме поиска к информации в отрыве от информационных потребностей, является практически неразрешимой задачей.

Основная идея адаптивного агрегирования информации заключается в сборе и хранении в информационном хранилище только той информации, которая соответствует информационным потребностям пользователей (существующих или потенциальных) [20]. Для этого предполагается, что по мере развития системы в ее информационное хранилище будут попадать актуальные документы из Интернет, соответствующие текущим запросам пользователей. Естественно, с ростом коли-

чества пользователей, объемы информационного хранилища (репозитария) будут также расти, что в некоторый момент потребует пересмотра его содержания по некоторым критериям, например, по времени в соответствии с формулой Бартона-Кеблера [21], или по содержанию, используя методы Text Mining.

Интернетика: теория сложных сетей, алгоритмы поиска в сложных сетевых структурах, элементы теории перколяции

В последнее время все большую популярность получает область дискретной математики, называемая теорией сложных сетей, изучающая характеристики сетей, учитывая не только их топологию, но и статистические феномены, распределение весов отдельных узлов и ребер, эффекты протекания и проводимости в таких сетях тока, жидкости, информации и т.д. Оказалось, что свойства многих реальных сетей существенно отличаются от свойств классических случайных графов. Самое подходящее название такой интеграции двух направлений — теорий информационного поиска и сложных сетей — Интернетика [16]. Во-первых, это направление является развитием информатики, и, что должно быть созвучно этому термину. Связь с теорией сложных сетей обуславливает наличие корня «нет», однако подразумевается, что исследования в рамках данного направления выйдут за рамки конкретной сети Интернет, анализ которой, безусловно, входит в сферу интернетики. Во-вторых, этот термин, хотя уже и встречается, но еще недостаточно устоялся. Известны, по меньшей мере, две трактовки термина «интернетика». В рамках первой интернетика рассматривается как прикладное научное направление, изучающее свойства и способы использования Интернет преимущественно в аспекте воздействия на социально-экономические процессы [22]. Эта трактовка несколько сужает область исследований (хотя и способствует популярности). Вторая трактовка заключается в том, что интернетика — это развитие информатики в направлении применения современных параллельных сетевых вычислений во всех областях науки, охватывая огромные ресурсы, распределенные в сетевой среде [23, 24]. Вторая трактовка понятия «интернетика», предполагающая использование методов точных наук, гораздо ближе авторам, чем первая.

Успешное продвижение в изучении современного информационного пространства невозможно без общих представлений о структуре и свойствах динамики сетевых информационных процессов, что в свою очередь требует выявления и учета их устойчивых закономерностей в рамках нового научного направления — «Интернетики».

Сегодня в Интернет существует доступная для экспериментов динамичная информационная база такого объема, который ранее даже трудно было представить. При этом оказалось, что многие задачи, возникающие при работе с сетевым информационным пространством, имеют немало общего, например, с задачами теоретической физики. Это обстоятельство открывает широкие перспективы применения мощного аппарата естественных наук.

Существует несколько актуальных задач исследования сложных сетей, среди которых можно выделить следующие основные:

— определение клик в сети. Клики — это подгруппы или кластеры, в которых узлы связаны между собой сильнее, чем с членами других клик;

— выделение компонентов (частей сети), которые связаны внутри и не связаны между собой;

— нахождение блоков и перемычек. Узел называется перемычкой, если при его исключении сеть распадается на несвязанные части;

— выделения группировок — групп эквивалентных узлов (какие имеют максимально похожие профили связей);

— выявление скрытых (латентных) связей [25];

— исключение незначительных (шумовых) связей;

— определение и учет динамики развития сети.

Заключение

Традиционно используемый математический аппарат и инструментальные средства информационного поиска сегодня уже не способны в полной мере удовлетворять потребности пользователей. Изначальная парадигма поисковых систем, сформированная несколько десятилетий тому назад, уже не отвечает реальной ситуации — объемам и динамике информационных потоков, сетевой топологии. Необходим поиск новых принципов, в рамках которых оказалось бы возможным проектирование качественно новых систем обработки больших и динамичных массивов данных.

Перспективы эффективного охвата информационного пространства будут зависеть как от создания и развития эффективной сетевой инфраструктуры, так и развития теоретических основ информатики. В этой связи одной из актуальнейших задач, стоящих перед исследователями различных специальностей, является построение адекватных моделей сетевого информационного пространства и информационного поиска, которые базируются на достижениях в областях лингвистики и информатики, строгом математическом инструментарии.

Предполагается, что должна быть создана теоретическая база для разработки автоматизированных систем мониторинга, адаптивного агрегирования и обобщения информационных потоков, построения и ведения информационных ресурсов сверхбольших объемов и разнообразной тематической направленности. Ожидаемые результаты позволят совместить в единой технологической цепочке мониторинг, информационный поиск, агрегирование информации с содержательным анализом данных, их обобщением, что повысит качество обработки сетевой информации, соответственно, эффективность информационно-аналитической поддержки научно-аналитической деятельности отечественных ученых и специалистов.

Данная статья подготовлена в рамках НИР по темам «Методы и средства анализа информационных потоков в компьютерных сетях для создания информационных ресурсов, ориентированных на решение аналитических задач», выполненной авторами в ИПРИ НАН Украины в 2009 г. и «Методы и средства мониторинга, адаптивного агрегирования и обобщения информации из глобальных компьютерных сетей для информационно-аналитической деятельности», которая ведется в настоящее время.

1. *Додонов О.Г.* Інформаційні потоки в глобальних комп'ютерних мережах / О.Г. Додонов, Д.В. Ланде, В.Г. Путятін. — К.: Наук. думка, 2009. — 295 с.

2. *Брайчевский С.М.* Современные информационные потоки: актуальная проблематика / С.М. Брайчевский, Д.В. Ландэ // Научно-техническая информация. — Сер. 1. — 2005. — Вып. 11. — С. 21–33.

3. *Ландэ Д.В.* Поиск знаний в Internet. Профессиональная работа / Д.В. Ландэ. — М.: Диалектика, 2005. — 272 с.

4. *Ландэ Д.В.* Дорожная карта сетевого поискового бизнеса / Д.В. Ландэ // Сети и бизнес. — 2009. — № 3. — С. 102–106.

5. *Ландэ Д.В.* Веб-пространство и материалы информационных агентств / Д.В. Ландэ, С.М. Брайчевский, А.Т. Дармохвал А.Т. // Компьютерная лингвистика и интеллектуальные технологии: Труды Международной конференции «Диалог». — Вып. 7(14). — М.: Изд-во РГГУ, 2008. — С. 303–305.

6. *Григорьев А.Н.* Адаптивный интерфейс уточнения запросов к системе контент-мониторинга InfoStream / А.Н. Григорьев, Д.В. Ландэ // Компьютерная лингвистика и интеллектуальные технологии: Труды Международного семинара «Диалог'2005» (Звенигород, 1–6 июня 2005 г.). — М.: Наука, 2005. — С. 109–111.

7. *Ландэ Д.В.* Подход к выявлению дублирования сообщений в новостных информационных потоках / Д.В. Ландэ, А.Т. Дармохвал, А.Ю. Морозов // Труды 8-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL'2006. — Суздаль, 2006. — С. 115–119.

8. *Зеленков Ю.Г.* Сравнительный анализ методов определения нечетких дубликатов для Web-документов / Ю.Г. Зеленков, И.В. Сегалович // Труды 9-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL'2007. — Переславль, 2007. — Т. 1. — С. 166–174.

9. *Никконен А.Ю.* Устранение избыточности и дублирования сюжетов новостных сообщений / А.Ю. Никконен // Интернет-Математика. Сборник работ участников конкурса. — Екатеринбург: Изд-во Урал. ун-та, 2007. — С. 157–167.

10. *Bourdaillet J.* Alignment of Noisy Unstructured Text Data / J. Bourdaillet // IJCAI-2007 Workshop on Analytics for Noisy Unstructured Text Data. — Hyderabad (India). — January 8, 2007. — P. 139–146.

11. *Salton G.* Vector Space Model for Automatic Indexing / G. Salton, A. Wong, C. Yang // Communications of the ACM. — 1975. — 18(11). — P. 613–620.

12. *Додонов А.Г.* Построение информационно-аналитической системы научно-исследовательского испытательного полигона / А.Г. Додонов, В.Г. Путятин, В.А. Валетчик // Управляющие системы и машины. — 2006. — № 4. — С. 3–14.

13. *Додонов А.Г.* Организация структуры Правительственной информационно-аналитической системы по вопросам чрезвычайных ситуаций / А.Г. Додонов, В.Г. Путятин, В.А. Валетчик // Электронное моделирование. — 2006. — Т. 28, № 3. — С. 61–82.

14. *Ландэ Д.В.* Основы интеграции информационных потоков / Д.В. Ландэ. — К.: Інжиніринг, 2006. — 240 с.

15. *Ландэ Д.В.* P2P — по секрету всему свету / Д.В. Ландэ // Сети и бизнес. — 2008. — № 2 (39). — С. 104–110.

16. *Ландэ Д.В.* Интернетика: Навигация в сложных сетях: модели и алгоритмы / Д.В. Ландэ, А.А. Снарский, И.В. Безсуднов. — М.: Либроком (Editorial URSS), 2009. — 264 с.

17. *Ландэ Д.В.* Структура новостного Web-пространства / Д.В. Ландэ // Научно-техническая информация. — Сер. 2. — 2006. — № 8. — С. 17–20.

18. *Массон Г.В.* Взаимосвязь системы личностных терминальных ценностей и типов межличностных отношений: дис. ... канд. психол. наук: 19.00.01. — Красноярск, 2004. — 146 с.

19. *Grishman R.* Information Extraction: Techniques and Challenges / R. Grishman // In Information Extraction (International Summer School SCIE-97). — Springer-Verlag, 1997. — P. 10–27.

20. *Bruton R.* The Half-Life of Some Scientific and Technical Literature / R. Bruton, R. Kebler R. // Am. document. — 1960. — N 1. — P. 18–22.

21. *Додонов А.Г.* Сетевые информационные потоки как содержательная составляющая информационно-аналитических систем / А.Г. Додонов, Д.В. Ландэ, В.В. Жигало // Реєстрація, зберігання і оброб.даних. — 2010. — Т. 12, № 1. — С. 39–48.

22. *Нехаев С.А.* Словарь прикладной интернетики [Электронный ресурс] / С.А. Нехаев, Н.В. Кривошеин, И.Л. Андреев, Я.С. Яскевич. — М.: Сетевой холдинг WEB-PLAN Group, 2001.

23. *Fox G.C.* From Computational Science to Internetics / G.C. Fox // Integration of Science with Computer Science, Mathematics and Computers in Simulation. — 2000. — N 54. — P. 295–306.

24. *Fox G.C.* Internetics: Technologies, Applications and Academic Fields / G.C. Fox // Invited Chapter in Book: Feynman and Computation, edited by A.J.G. Hey, Perseus Books (1999). Technical Report SCCS-813. — Syracuse University, NPAC, Syracuse, NY. — February 1998.

25. *Clauset A.* Hierarchical Structure and the Prediction of Missing Links in Networks / A. Clauset, C. Moore, M.E.G. Newman // Nature. — 2000. — Vol. 453. — P. 98–101.

Поступила в редакцию 21.06.2010