УДК 004.9

А. М. Федотов, В. Б. Барахнин

Институт вычислительных технологий СО РАН
пр. Акад. Лаврентьева, 6, Новосибирск, 630090, Россия

Новосибирский государственный университет
ул. Пирогова, 2, Новосибирск, 630090, Россия

E-mail: bar@ict.nsc.ru, fedotov@nsu.ru

ПРОБЛЕМЫ ПОИСКА ИНФОРМАЦИИ: ИСТОРИЯ И ТЕХНОЛОГИИ*

Статья посвящена обсуждению проблем поиска информации в современной информационной среде, историческим подходам, технологическим задачам и алгоритмам.

Ключевые слова: информационный поиск, Интернет

Введение

Проблема поиска информации – одна из вечных проблем человеческого сообщества. На протяжении своего многотысячелетнего развития его представители неустанно находятся в поиске того, где находится что-либо*: пищи, жилища, пастбищ, дорог, сокровищ* и т. п. Обобщая задачи поиска, можно сказать, что человечество постоянно находится в поиске *знаний*, а в частности «информации о том, где лежат сокровища». Великий аргентинский писатель Хорхе Луис Борхес [1] в своем эссе «Четыре цикла» писал, что в мировой литературе вечными являются четыре темы:

1. Падение города.
2. Возвращение героя.
3. Поиск.
4. Самопожертвование бога.

Нетрудно заметить, что наиболее часто встречающейся как в литературе, так и в реальности является третья тема – *поиск*, ибо четвертая тема выходит за рамки обычного человеческого опыта, а две первые проявляются лишь в «минуты мира роковые».

С появлением новой экономической категории [2], какой являются информационные ресурсы, проблема поиска перекочевала и в эту область. Человечество все больше начинает использовать для поиска необходимых знаний информационные ресурсы. Чтобы решить проблему доступа к информации, человечество создало библиотеки – как универсальную систему хранения «знаний», их систематизации и каталогизации.

Ситуация кардинально изменяется по мере освоения (точнее – создания) человеческой цивилизацией пространства «информационного». Первыми островами информационного пространства цивилизации стали общественные библиотеки [3], крупнейшие из которых (Библиотека Британского музея, Национальная библиотека в Париже, Библиотека конгресса США, Российская государственная библиотека и др.) уже к началу XX в. располагали собраниями в миллионы томов.

* Работа выполнена при частичной финансовой поддержке РФФИ (проекты № 07-07-00271, 08-07-00229, 09-07-00277), президентской программы «Ведущие научные школы РФ» (грант № НШ-931.2008.9) и интеграционных проектов СО РАН.

[1] Хорхе Франсиско Исидоро Луис Борхес Асеведо – Jorge Francisco Isidoro Luis Borges Acevedo.

[2] Информация и информационные ресурсы существовали всегда, но эти ресурсы из-за своей специфичности не рассматривались ранее как отдельная экономическая категория, несмотря на то, информация всегда использовалась людьми для управления и решения насущных задач.

[3] Здесь мы не будем говорить о крупнейших библиотеках древности, поскольку в них проблемы поиска не были столь актуальны.

Долгое время одним из мощных инструментов поиска информации в книжных хранилищах был непосредственный доступ читателей к книгам, когда они, затрачивая большое личное время, могли свободно рыться в библиотеке. Это и понятно, поскольку человека, нуждающегося в научной информации (в знаниях), интересует прежде всего не сама книга как таковая, а только некоторый ее фрагмент, содержащий требуемые ему знания. Причем сам он часто не в состоянии объяснить, как эти знания могут быть связаны с названием книги или ее автором.

Накопление книг привело к парадоксальному результату, связанному с отделением книжных хранилищ от широкого круга читателей. Универсальный инструмент поиска знаний, основанный на прямом доступе к информации, стал доступен только избранным. Основная же масса жаждущих знаний стала довольствоваться только поиском в каталоге, который в принципе не мог удовлетворить возникающие информационные потребности. Для решения проблемы доступа читателей к информации были предприняты попытки классификации и систематизации информации – стали создаваться специализированные книжные залы, куда источники информации отбирались исходя из каких-то (не всегда очень ясных) критериев.

С одной стороны, как отметил известный историк и социолог науки Д. де Солла Прайс [4] [Price D. J. de Solla 1966], начиная с середины XVIII в. любой достаточно большой сегмент науки в нормальных условиях растет экспоненциально, т. е. любые параметры науки, включая объем накопленной информации, за определенный промежуток времени удваиваются (закон экспоненциального роста науки). С другой стороны, в указанный период времени, происходит увеличения числа людей, нуждающихся в научной информации. Речь идет не только о научных работниках (численность которых тоже подчиняется закону экспоненциального роста), но и о представителях многих других профессий умственного труда: инженерах, агрономах, врачах, управленцах и т. п.

По мере накопления книг, а стало быть, и содержащейся в них информации, возможности традиционных методов поиска: с использованием алфавитного каталога (поиск книги по известному имени автора) и систематического каталога (поиск книги или класса книг по определенному предмету), – перестали удовлетворять читателей, прежде всего научных работников, информационные потребности которых в процессе научного поиска характеризуются невысокой четкостью осознания и выражения (см., например: [Арский и др., 1996]).

Современные информационные технологии предоставляют исследователю мощный аппарат для «манипулирования данными», а не информацией. Данные, переведенные в электронную форму, приобретают новое качество, обеспечивая им более широкое распространение и эффективное использование. На первый взгляд, может сложиться впечатление, что развитие информационных технологий уже само по себе способно вывести работу с научной информацией на качественно новый уровень, но, к сожалению, это совсем не так. Современные информационные технологии пока не могут предоставить адекватный аппарат для оперирования с «информацией» и информационными ресурсами [Черняк, 2004].

Однако сами по себе данные (как набор битов) не представляют никакой информационной ценности без соответствующих описаний или моделей. Применение информационных технологий должно основываться на использовании различных моделей (феноменологических, информационных, математических и др.). Как неоднократно отмечал А. А. Ляпунов (см., например: [Ляпунов, 1980]): «нет модели – нет информации».

Для возможности продуктивной работы нужны данные, превращенные в «информацию», представленную в виде «знаний» – «адекватного отражения действительности в сознании человека в виде представлений, понятии, суждений теорий».

Существующую проблему отбора информации уже дано пытаются решить путем создания универсальных или специализированных информационно-поисковых систем. В результате опережающего развития технологий поиска по сравнению с методиками работы с семантической информацией образовался заметный разрыв между техникой работы с данными (поиском) и способностью работать с содержанием, заложенным в этих данных. Опираясь на интуицию, эксперты приходят к выводу о порочности нынешней ситуации, но о каком-либо серьезном переосмыслении проблем извлечения из данных информации пока речь не идет.

[4] Дерек Де Солла Прайс – Derek J. de Solla Price.

Предыстория

Как мы видим, что проблема поиска – доступа к информации является одной из серьезных проблем, с которой столкнулось современное «информационное общество».

По всей видимости, впервые возникшую проблему наиболее четко осознал бельгийский социолог Поль Отле [5], который в конце XIX в. предложил дополнить науку [6] (library science), ведавшую научно-технической информацией и традиционное библиотековедение совершенно новым методом, названным им «Документацией».

«Цели Документации состоят в том, чтобы суметь предложить документированные ответы на запросы по любому предмету в любой области знания: 1) универсальные по содержанию; 2) точные и истинные; 3) полные; 4) оперативные; 5) отражающие последние данные; 6) доступные; 7) заранее собранные и готовые к передаче; 8) предоставленные как можно большему числу людей» (см.: [Отле, 2004. С. 190; Otlet, 1934]).

Суть метода *Документации* заключалась в том, что содержание книги (отчуждаемое от автора) заносится на карточку, причем совокупность карточек можно упорядочивать так, чтобы при этом отражались предметные связи. Поль Отле предвидел революционное развитие технологий работы с информацией, вплоть до ее мультимедийного представления и удаленного доступа к банкам данных: «…человеческое знание позволит создать оборудование, действующее на расстоянии, в котором соединятся радио, рентгеновские лучи, кинематограф и микроскопическая фотография. Все предметы Вселенной, все предметы, созданные Человеком, будут регистрироваться на расстоянии с момента их создания. Тем самым будет создан движущийся образ мира – его память, его подлинная копия. Любой человек сможет прочесть отрывок, спроецированный на его личный экран» (см.: [Отле, 2004. С. 16]).

Идеи Поля Отле не были восприняты тогдашними информационным (библиотечным) сообществом, в частности потому, что они совершенно не были подкреплены техническим обеспечением: информационные работники и библиотекари той эпохи располагали лишь пишущими машинками, фотоаппаратами и карточными каталогами. Появление после Первой мировой войны устройств обработки перфокарт (точнее, их простейшей разновидности – перфокарт с краевой перфорацией) также не стало принципиальным технологическим прорывом, поскольку даже спустя 40 лет, в 1960-е гг., подобные устройства могли обрабатывать сравнительно небольшие (до 30 тыс.) массивы документов (см.: [Михайлов и др., 1968. С. 549]).

Проблема нарастающих объемов информации, грозивших захлестнуть читателей, продолжала волновать исследователей. В 1941 г. упомянутый выше Х. Л. Борхес создает свою знаменитую притчу «Вавилонская библиотека». В этой притче Вселенная представляется в виде Библиотеки, беспредельной и всеобъемлющей, на полках которой «можно обнаружить все возможные комбинации двадцати с чем-то орфографических знаков (число их, хотя и огромно, не бесконечно) или все, что поддается выражению – на всех языках». Философский смысл притчи, конечно же, гораздо глубже проблемы [7] информационного поиска, но исходный образ взят автором из повседневной реальности. Трудно удержаться, чтобы не привести хотя бы краткие выдержки из притчи, соответствующие тематике статьи.

«Когда было провозглашено, что Библиотека объемлет все книги, первым ощущением была безудержная радость. Каждый чувствовал себя владельцем тайного и нетронутого сокровища. Не было проблемы – личной или мировой, для которой не нашлось бы убедительного решения… Вселенная обрела смысл, вселенная стала внезапно огромной, как надежда. В это время много говорилось об Оправданиях: книгах апологии и пророчеств, которые навсегда оправдывали деяния каждого человека во вселенной и хранили чудесные тайны его будущего. Тысячи жаждущих покинули родные шестигранники и устремились вверх по лестницам, гонимые напрасным желанием найти свое оправдание…, но те, кто пустился на по-

[5] Поль Отле – Paul Otlet.

[6] Термин «информатика» принадлежал когда-то скромной науке, ведавшей именно информацией, в основном научно-технической. Термин «информатика» (фр. informatique) родился в 1960 г., условно происходит от французских слов information (информация) и automatique (автоматизация) и дословно означает «информационная автоматизация».

[7] Борхес был профессиональным библиотекарем (библиографом) и даже одно время занимал пост директора Национальной библиотеки Аргентины.

иски, забыли, что для человека вероятность найти свое Оправдание или какой-то его искаженный вариант равна нулю…

На смену надеждам, естественно, пришло безысходное отчаяние. Мысль, что на какой-то полке в каком-то шестиграннике скрываются драгоценные книги и что эти книги недосягаемы, оказалась почти невыносимой. Одна богохульная секта призывала всех бросить поиски и заняться перетасовкой букв и знаков, пока не создадутся благодаря невероятной случайности канонические книги: Другие, напротив, полагали, что прежде всего следует уничтожить бесполезные книги …

Известно и другое суеверие того времени: Человек Книги. На некоей полке в некоем шестиграннике (полагали люди) стоит книга, содержащая суть и краткое изложение всех остальных: некий библиотекарь прочел ее и стал подобен Богу. В языке этих мест можно заметить следы культа этого работника отдаленных времен. Многие предпринимали паломничество с целью найти Его. В течение века шли безрезультатные поиски. Как определить таинственный священный шестигранник, в котором Он обитает? Кем-то был предложен регрессивный метод: чтобы обнаружить книгу А, следует предварительно обратиться к книге В, которая укажет место А; чтобы разыскать книгу В, следует предварительно справиться в книге С, и так до бесконечности …»

Движущей силой произошедшей в середине XX в. «информационной революции» стали не хранители информации – библиотечные работники, а ее потребители – ученые и инженеры. В 1931 г. в Германии была создана Статистическая машина Эммануэля Гольдберга [8] [Черняк, 2004], обеспечивавшая чтение специальным образом подготовленной микропленки, на которой хранился массив документов. Особенность организации хранения информации заключалась в том, что на пленку вместе с микрофильмированным документом заносилось описание этого документа, закодированное посредством перфорации. Поиск документа осуществляется путем сравнения запроса (также закодированного) с перфорацией пленки. Машину Гольдберга отличало высокое качество механики и оптики: пользователь имел возможность просматривать за час более 100 000 кадров 35-миллиметровой пленки. Статистическая машина Гольдберга была, по-видимому, первым действующим инструментом, позволяющим автоматизировать поиск в больших массивах данных по их разметке. Кстати сказать, по мнению некоторых исследователей, на идеи Эммануэля Гольдберга опирался Вэннивер Буш [9], автор знаменитой статьи «Пока мы мыслим» («As We May Think») [Bush, 1945], фактически написанной в 1939 г., в которой сформулирована идея гипертекста и предсказано появление персонального устройства, хранящего информацию и автоматизирующего процесс ее поиска. Вот как выглядит одна из его идей:

«Обсудим устройство персонального назначения. Пусть оно называется Memex и представляет собой что-то вроде автоматизированного архива или библиотеки. Memex хранит для своего хозяина все нужные книги, записи, корреспонденцию. Прибор автоматизирован до такой степени, что дает ответы на вопросы, заданные в простой форме, – т. е. очень гибок в общении.

Скорость ответов высока и не заставляет ждать. Имеется графический экран, клавиатура и кнопки управления. Когда пользователь ищет нужную книгу, он должен ввести ее мнемонический код и нажать нужную для поиска кнопку. Перед ним на экране появится первая страница. Должна быть возможность листать книгу в любом направлении. Можно будет остановиться на выбранной странице, а потом пойти по ссылке и найти следующий интересующий материал. При этом всегда можно вернуться к предыдущей странице или одновременно рассматривать несколько страниц.

Появятся энциклопедии с готовыми ссылками для связывания информации и быстрого поиска. Их можно будет загружать в Memex и искать все, что нужно».

Нередко в литературе можно встретить высказывания, что В. Буш предсказал идею персонального компьютера, но так говорить не совсем правильно, ибо фактическое время написания статьи «As We May Think» относится к тому периоду, когда под руководством В. Буша

[8] Эммануэль Гольдберг – Emanuel Goldberg – немецкий инженер, выходец из России.
[9] Вэннивер Буш – Vannevar Bush.

в Массачусетском технологическом институте был создан действующий макет микрофильмового селектора «Мемекс» [Михайлов и др., 1968].

Если же говорить о поисковых устройствах той эпохи, основанных не на аналоговом, а на цифровом представлении информации (как раз и используемом в современных компьютерах), то следует отметить реализованную на суперпозиционных перфокартах систему поиска патентов, которую в 1939 г. создал У. Баттен для британского концерна «Imperial chemical industries, Ltd». Ее алгоритм работы был основан на координатном индексировании – представлении содержания документа при помощи списка содержащихся в нем ключевых слов. Эта идея получила дальнейшее развитие в работах американского математика Кельвина Муэрса [10], создавшего и запатентовавшего в 1947 г. систему механизированного поиска документов, работавшую на особых картах с вырезами вдоль краев (так называемых «Zato-картах»).

В основе системы также лежал метод координатного индексирования. Именно К. Муэрс стал основоположником научного подхода к информационному поиску, введя в 1950 г. термины «информационный поиск», «информационно-поисковая система», «информационно-поисковый язык», «поисковый образ», «дескриптор», «дескрипторный словарь» и др. С этого времени началось бурное развитие информатики как науки о структуре и свойствах семантической информации (прежде всего научной). Важное место в этой науке занимали вопросы информационного поиска, в процессе выполнения которого, собственно говоря, и происходит непосредственное удовлетворение информационных потребностей пользователя. Обобщение накопленных результатов было проведено в монографии сотрудников Всесоюзного института научной и технической информации (ВИНИТИ) [Там же], описавших методологические основы теоретической информатики.

Возможности практической реализации алгоритмов информационного поиска резко расширились, когда в середине 1960-х – начале 1970-х гг. вместо механических устройств стали достаточно широко применять электронно-вычислительные машины третьего, а затем и четвертого поколений, на базе которых создавались автоматизированные системы сбора, анализа, классификации, хранения, передачи на расстояние, поиска и выдачи информации. В частности, исследовательская группа под руководством профессора Гарвардского университета Дж. Солтона [11] разработала систему анализа и извлечения текста SMART (Salton's Magic Automatic Retriever of Text), в которой были впервые реализованы многие базовые принципы современных поисковых систем. Теоретическое описание и осмысление этих принципов было проведено Дж. Солтоном в монографии [1979], причем особый акцент в ней был сделан на изложении новых подходов к вопросам классификации документов и запросов, анализ содержания, интерактивного поиска и выдачи информации. Эта книга и до сих пор не потеряла своей актуальности.

Технологической основой создания подобных информационно-поисковых систем было использование так называемых мэйнфреймов – многопользовательских централизованных вычислительных систем, в которых массивы данных и программы их обработки располагались на мощной центральной ЭВМ, а пользовательский доступ осуществлялся посредством алфавитно-цифровых терминалов (дисплеев), работающих под управлением машин-сателлитов. Бытует мнение, что информационно-поисковые системы того времени не получили должного развития из-за недостаточной мощности и памяти тогдашних ЭВМ, так и с отсутствием качественных каналов связи (особенно дальней). Здесь проблемы были несколько другие. Во-первых, отсутствие универсальных сетевых протоколов сильно ограничивало удаленный доступ к таким системам. Во-вторых, большая загрузка вычислительными задачами не позволяла организовать работу таких систем в круглосуточном режиме. Все это придавало информационно-поисковым системам преимущественно локальный характер.

Несмотря на это, в информационных системах того времени был собран и систематизирован колоссальный по тем временам объем информации. Например, в Новосибирском ВЦ СО РАН на машинах типа БЭСМ-6 хранились вся подписках реферативных журналов ВИНИТИ, библиографические описания изданий, поступающих в ГПНТБ, и большое количество науч-

[10] Кельвин Муэрс – Calvin Northrup Mooers.
[11] Джерард Солтон – Gerard Salton.

но-технической документации. Основные проблемы связанные с ее использованием – это отсутствие интерактивной работы, поскольку, как правило, запрос посылался с терминала, а ответ приходил в виде «километровой» распечатки на АЦПУ. И это была жизненная необходимость, поскольку анализировать ответ за дисплеем не представлялось никакой возможности. Ну а вторая проблема была связана с визуализацией материала – практически отсутствовало программное обеспечение, позволявшее просматривать информацию в близком к печатному изданию виде.

В 1980-е гг. мэйнфреймы стали постепенно вытесняться персональными компьютерами, которые позволяли обрабатывать информацию непосредственно на рабочем месте, без связи с центральным процессором, а кроме того, обладали достаточно мощными (по тем временам) средствами визуализации информации. Это привело к существенному снижению интереса к созданию централизованных информационных систем и, как следствие, к приостановке фундаментальных научных исследований в области информационного поиска, которые возобновились лишь с появлением сети Интернет, приведшим к распределенному хранению информации.

Принципы организации информационно-справочных систем

Как уже отмечалось, созданные в трудах К. Муэрса и Дж. Солтона фундаментальные основы поиска информации являются актуальными и по сей день. Однако здесь есть небольшой нюанс в их использовании. «Классики» называли такие системы Information Retrieval System (IRS). В 1950–1970 гг. англоязычный термин Information Retrieval (IR) переводили на русский язык как «информационный поиск», а соответственно, системы этого класса называли информационно-поисковыми системами. В этих системах использовались ручные процедуры индексирования документов, создания тезаурусов и дескрипторов. Но, что чрезвычайно важно, эти системы предназначались для *выделения* информации (именно информации и именно выделения) из разных документов. «Выделение» – это более точное значение слова retrieval. Сейчас в энциклопедиях IR определяется как искусство и наука поиска информации в документах и поиска собственно документов и описывающих документы метаданных в базах данных (в том числе сетевых). Подмножеством IR является выделение информации в тексте (Text Retrieval, TR) и выделение информации в документах (Document Retrieval, DR).

Мы напоминаем об этом, чтобы подчеркнуть различие между поиском как автоматизированной процедурой и выделением требуемой информации в найденных документах. Суть различий состоит в следующем.

- Выделение информации – это деятельность человека, использующего поисковую машину. Она является интерактивной, итерационной и связана с другими видами интеллектуальной деятельности человека.
- Читатель ищет не документы как таковые, а содержащуюся в них информацию для каких-то собственных целей (обучения, принятия решений и др.).
- Читатель нуждается в доступе к разным источникам данных, чтобы получить всеобъемлющее представление об объекте поиска.
- Какими бы совершенными ни были аппаратное и программное обеспечение, используемые человеком, они остаются инструментами, а интеллект является атрибутом Читателя.

Наиболее радикальный этап «информационный революции» начался в 1990-е гг. Он был связан с по-настоящему массовым распространением мощных и недорогих персональных компьютеров, которые могли быть подключены в созданную всемирную компьютерную сеть Интернет. Именно сеть Интернет, отличающаяся от печатных изданий оперативностью размещения и доставки информации практически любого характера, а от классических электронных СМИ – возможностью передачи печатного текста, делает все более реальной перспективу создания единого информационного пространства человеческой цивилизации.

В настоящее время интернет является главным источником электронных документов. Количество документов в сети поддается лишь косвенным, притом явно заниженным оценкам. Так, по состоянию на начало августа 2005 г. число документов, проиндексированных поисковой системой Yahoo, превысило 20 млрд, из них 19,2 млрд – текстовые документы,

1,6 млрд – изображения и около 50 млн – аудио- и видеофайлы [12]. При этом, разумеется, нельзя утверждать, что Yahoo индексирует все интернет-документы.

Однако такое обилие потенциально доступных документов сделало особенно актуальной задачу предоставления пользователям сети адекватных средств информационного поиска, без которых интернет мог бы превратиться в реальное воплощение «Вавилонской библиотеки». Говоря о средствах информационного поиска в сети Интернет, обычно подразумевают *поисковые системы* – веб-сайты, предоставляющие возможность поиска информации по всему Интернету (по крайней мере, по всем www-страницам). Такие системы известны всем пользователям интернета: это Google, Yahoo, MSN и др. (из числа отечественных разработок наиболее популярны Yandex, Rambler и Mail.ru). Однако для поиска документов, относящихся к той или иной предметной области, пользователи интернета нередко обращаются к *тематическим каталогам интернет-ресурсов* – структурированным наборам ссылок на документы соответствующей тематики.

Чтобы описать принципы работы средств информационного поиска, необходимо, прежде всего, уточнить соответствующую терминологию. Основные термины и определения в области поиска и распространения информации с помощью автоматизированных информационных систем, а также информационно-поисковых языков регламентированы официальными документами Российской Федерации (действующими и в большинстве других стран СНГ): государственными стандартами ГОСТ 7.73-96 «Поиск и распространение информации» и ГОСТ 7.74–96 «Информационно-поисковые языки».

Итак, информационно-поисковая система (ИПС) представляет собой совокупность справочно-информационного фонда и технических средств информационного поиска в нем. В свою очередь, справочно-информационный фонд (СИФ) – это совокупность информационных массивов (т. е. упорядоченных совокупностей документов, фактов или сведений о них) и связанного с ними справочно-поискового аппарата (т. е. данных об адресах хранения документов с определенными поисковыми образами документа). Наконец, поисковый образ документа – это текст, состоящий из лексических единиц информационно-поискового языка (т. е. специального формализованного искусственного языка), выражающий основное смысловое содержание документа и предназначенный для реализации информационного поиска. Процесс выражения содержания документа на информационно-поисковом языке называется индексированием.

Заметим, что под содержанием документа в данном контексте обычно подразумевают не только более или менее краткое изложение того, о чем повествует документ, но и его «библиографические характеристики»: название документа, фамилии его авторов, выходные данные и т. п. Совокупность извлекаемых в процессе индексации характеристик документа вместе с формальным описанием структуры этих характеристик обычно называют *метаданными*. Более формально, метаданные – это структурированные данные, представляющие собой характеристики описываемых сущностей для целей их идентификации, поиска, оценки, управления ими [Task Force…, 1999].

Структурирование данных призвано облегчить поиск документов, ибо одно и то же слово (например, «Пушкин») может входить в список авторов документа, в его заглавие, в аннотацию или даже в выходные данные (город Пушкин в Ленинградской области как место издания документа). Эти случаи могут быть разграничены именно благодаря структурированию метаданных.

Нетрудно понять, что документ становится доступным для поиска с помощью той или иной информационно-поисковой системы, если его метаописание (т. е. совокупность метаданных) попадает в справочно-информационный фонд этой системы. Но каким образом осуществляются поиск и индексация интернет-документов, заносимых в СИФ? Поисковые системы общего назначения используют поисковые роботы (их английское название – «crawler», т. е. «ползун»), которые последовательно просматривают интернет-документы, переходя от одного к другому посредством гиперссылок, и извлекают их метаданные. Разумеется, поисковые роботы периодически просматривают документы, уже занесенные в СИФ информационной системы, чтобы установить, существуют ли они в настоящее время и не

[12] *Mayer T.* Our Blog is Growing Up – And So Has Our Index. http://www.ysearchblog.com/archives/000172.html

претерпели ли они каких-либо существенных изменений. При составлении тематических каталогов интернет-ресурсов также зачастую используются поисковые роботы, которые, однако, собирают данные о документах лишь с сайтов соответствующей тематики. Сетевые имена таких сайтов, как правило, указываются экспертами в данной предметной области, при этом допускается и непосредственное занесение экспертами сведений об отдельных интернет-документах. Наконец, некоторые специализированные информационно-поисковые системы создаются исключительно вручную, при этом размер их поисковых массивов может быть весьма внушителен. Так, очень популярная в среде математиков база данных журнала «Zentralblatt MATH» содержит почти 3 млн записей – библиографических сведений (включая довольно подробные аннотации) о математических публикациях, вышедших в свет за последние полтора века. Эти сведения заносятся в базу данных учеными-математиками из разных стран, реферирующими публикации по своей специальности, причем каждой записи соответствует динамически формируемый интернет-документ.

Но все-таки справочно-информационные фонды большинства информационно-поисковых систем, работающих с интернет-документами, пополняются не вручную, а с помощью тех или иных программ, автоматизирующих поиск и индексацию документов. И здесь-то, в процессе индексации документа, проявляется основная проблема использования таких программ: автоматическое структурирование метаданных оказывается весьма непростой задачей. Чтобы убедиться в этом, достаточно просмотреть небольшое число интернет-документов, например, научной тематики. Можно легко увидеть, что в некоторых случаях фамилии авторов пишутся перед названием документа, а в некоторых, наоборот, после названия. Каким образом программа должна определять, что именно заносить в поле «авторы» данного документа, а что – в поле «название»? Заметим, что простейшие варианты решения этой проблемы (типа «дополнить индексирующую программу словарем фамилий») оказываются малоэффективными. И дело не только в необходимости огромного (и не существующего на практике) объединенного словаря фамилий разных наций с вариантами транскрипций на других языках. Проблема состоит еще и в том, что многие фамилии (особенно в языках со слабовыраженным изменением словоформ при помощи окончаний) с «обычными» словами языка. Кроме того, фамилия может являться названием документа, например книги или статьи биографического характера.

Наличие указанных проблем привело к тому, что обычной практикой универсальных поисковых систем является представление поискового образа документа в виде неструктурированного набора *ключевых слов* – информативных слов, приведенных к стандартной лексикографической форме. *Информативными словами*, согласно ГОСТу 7.74-96, называются слова, словосочетания или специальные обозначения в тексте документа (или запроса), выражающие понятия, существенные для передачи содержания документа. Конкретные критерии включения слова или словосочетания к множеству информативных слов зависят от вида ИПС. Так, в универсальных поисковых системах в качестве информативных рассматриваются практически все слова, включая служебные. Напротив, в специализированных информационно-поисковых системах, для которых набор ключевых слов – один из компонентов структуры метаданных документа, множество информативных слов обычно строится на основе предметного указателя соответствующей предметной области (содержащего наряду с одиночными словами и весьма сложные словосочетания), в то время как слова, относящиеся к «общеупотребительной» лексике, в число информативных не включаются.

Поскольку совершенно очевидны преимущества структурированного описания документа перед неструктурированным (о чем уже говорилось выше), постольку организациями, пытающимися выступать в качестве «законодателе мод» в сети Интернет, прежде всего консорциумом W3C, неоднократно предпринимались попытки предоставить создателям интернет-документов возможность *явно* указывать значения основных элементов метаданных документа, что позволило бы значительно повысить эффективность функционирования поисковых роботов. Так, еще в середине 1990-х гг. в спецификации языка гипертекстовой разметки документов HTML было четко прописано, что каждый html-документ обязан иметь ровно один элемент TITLE («название») в поле HEAD («заголовок»). Более того, в описании языка HTML появился элемент META, предназначенный для записи парных элементов

NAME:CONTENT («название:значение»), описывающих свойства данного документа: фамилия автора, список ключевых слов и т. п.

Заметим, однако, что спецификация языка HTML не предусматривала каких-либо *конкретных* названий для обозначения элементов, содержащих информацию о фамилии автора, ключевых словах и пр. Ввиду этого даже при наличии в индексируемом документе элементов META задача автоматического определения его структуры оставалась трудноразрешимой. Наиболее известным подходом к ее решению стал предложенный в 1995 г. на семинаре, проводившемся Национальным центром суперкомпьютерных приложений (NSCA) в городе Дублин (штат Огайо, США), базовый набор из 15 полей метаданных, предназначенный для описания ресурсов, публикуемых в интернете. В этот набор вошли такие общие свойства документов, как название, дата публикации, автор, издатель, владелец. Таким образом, в любом документе должно было существовать ядро метаданных, о которых заранее известно, как их следует интерпретировать. Эти предложения были опубликованы под рабочим названием Dublin Core metadata, которые впоследствии стали фундаментом проекта Dublin Core Metadata Initiative [13].

Названные идеи получили дальнейшее развитие в проекте Semantic Web, суть которого заключается в создании сети документов, содержащих метаданные «исходных» документов сети Интернет и существующей параллельно с ними. Эта «параллельная» сеть предназначена специально для построения поисковыми роботами (и другими интеллектуальными агентами) однозначных логических заключений о свойствах «исходных» документов. Основные принципы создания Semantic Web (до практической реализации которой, впрочем, еще очень далеко) основаны на повсеместном использовании, во-первых, универсальных идентификаторов ресурсов (URI) посредством расширения этого понятия на объекты, недоступные для скачивания из Интернета (персоны, географические сущности и т. п.), а во-вторых – онтологий (т. е. формальных моделей описания тех или иных предметных областей) и языков описания метаданных.

К сожалению, ни один из перечисленных подходов не стал по-настоящему широко распространенным. В этом без труда можно убедиться, просмотрев произвольный набор интернет-документов. Почти наверняка в большинстве из них будут отсутствовать элементы META, содержащие фамилии авторов, список ключевых слов и т. п. Причины сложившейся ситуации широко обсуждаются в интернет-сообществе, но, несомненно, к числу основных причин относится «человеческий фактор».

Во-первых, ввиду широкой распространенности интернет-технологий теоретическая подготовка многих создателей интернет-ресурсов оставляет желать лучшего, и они зачастую просто не знают о назначении элемента META в языке HTML. Во-вторых, явное указание значений метаданных – процесс весьма трудоемкий, поэтому даже те создатели ресурсов, которые знают о технологии метаданных, не всегда считают нужным тратить время и силы на работу с ними, тем более что разработчики универсальных поисковых систем, исходя из описанной ситуации, не слишком-то полагаются на возможность автоматического получения структурированного поискового образа индексируемого документа, ибо процент документов, подробно описанных создателями, весьма невелик. В итоге складывается своеобразный порочный круг, который в ближайшее время вряд ли будет разорван.

В несколько лучшем положении находятся создатели тематических каталогов интернет-ресурсов, поскольку количество организаций, работающих в той или иной области человеческой деятельности, а также веб-сайтов, публикующих действительно ценную и/или новую информацию соответствующей тематики, как правило, довольно невелико. Важно отметить, что реальные технологии создания подавляющего большинства сайтов таковы, что однородные документы с одного сайта имеют практически одинаковую html-разметку. При этом неважно, генерируются ли документы динамически (в этом случае однородность разметки – естественное следствие работы соответствующей программы) или же они создаются вручную посредством создания копии уже имеющегося документа с последующей заменой текста (что также сохраняет разметку). Данное обстоятельство позволяет автоматизировать процесс индексации метаданных интернет-документа посредством указания шаблона документов то-

[13] http://dublincore.org/

го или иного сайта, т. е. явному указанию команд (тэгов) языка HTML, обрамляющих основные характеристики документа: авторы, название, ключевые слова, аннотация, коды того или иного классификатора и т. п. [Барахнин, Федотов, 2008].

Составление поисковых предписаний

Из предыдущего пункта мы получили некоторое представление о том, как устроен справочно-информационный фонд ИПС. Чтобы сделать запрос, мы должны, прежде всего, составить поисковый образ запроса, т. е. его формальное представление в терминах информационно-поискового языка. После этого составляется *поисковое предписание*, включающее поисковый образ запроса и указания о логических операциях, подлежащих выполнению в процессе информационного поиска. ИПС сравнивает поисковое предписание с хранящимися в ее справочно-поисковом аппарате поисковыми образами документов (при этом в большинстве поисковых систем ключевые слова по умолчанию приводятся к стандартной лексикографической форме) и выдает сведения: адреса хранения и, как правило, краткие описания, – о документах, поисковые образы которых соответствуют (т. е. фактически не противоречат) поисковому предписанию.

Например, поисковое предписание для ИПС интернет-магазина, торгующего мужскими костюмами, может выглядеть примерно так:

(рост = 176) и (размер = 104) и ((цвет = ’черный’) или (цвет = ’темно-синий’)) и (страна-производитель = не ’Китай’) и (цена < 7000 руб.)

При этом, коль скоро не указаны значения таких элементов метаданных, как материал и тип костюма (пара или тройка), то подразумевается, что пользователя устраивают любые значения этих элементов метаданных.

Простейшая формальная модель с использованием структурированных метаданных документов выглядит следующим образом. Пусть в справочно-поисковом аппарате ИПС хранится информация о документах d_i. При этом любой документ d_i представляется как $d_i = < m_i^{j,k} >$, где $m_i^{j,k}$ – значения элементов метаданных M^j, k – количество значений (с учетом повторений) соответствующего элемента метаданных в описании документа. Рассмотрим подмножество метаданных M_C, определяющее набор классификационных признаков документов, используемых для составления поискового предписания (с учетом заданных логических операций). Для фиксированного элемента метаданных M_j, где $M_j \subset M_C$, множество документов разбивается на классы эквивалентности, соответствующие различным значениям этого элемента метаданных.

Будем считать два документа *толерантными*, если у них совпадает значение хотя бы одного из элементов метаданных, входящих в M_C (напомним, что толерантность – отношение, которое обладает свойствами рефлексивности и симметричности, но, вообще говоря, может не обладать, в отличие от отношения эквивалентности, свойством транзитивности). Каждое такое значение порождает класс толерантности [Шрейдер, 1971].

Рассмотрим всевозможные сочетания значений элементов метаданных, входящих в M_C. Множества документов, обладающие одинаковым набором значений, суть ядра толерантности, которые служат классами эквивалентности на множестве документов.

Таким образом, поисковое предписание, содержащее подмножества метаданных, определяющего набор классификационных признаков, и сочетаний значений этих метаданных при помощи логических операций, определяет конкретное ядро толерантности на множестве документов, которое и выдается пользователю в качестве ответа на его информационный запрос.

К сожалению, в ИПС общего назначения поисковые образы документов, как уже отмечалось в предыдущем пункте, структурированы весьма слабо. Обычно пользователь таких систем имеет возможность включить в поисковый образ запроса (точнее, в ту его часть, которую описывает *содержание* требуемого документа) лишь ключевые слова или словосочетания, указав при этом, где именно они должны содержаться: в заголовка веб-страницы или в ее тексте. Остальные поля в форме поискового запроса касаются языка документа, региона расположения сервера размещения документа, формата файла, структуры его url-адреса и т. п., т. е. не имеют непосредственного отношения к содержанию документа.

Впрочем, построение более или менее сложного поискового предписания способно вызвать затруднение у большинства рядовых пользователей, даже если им предоставлен удобный интерфейс, не требующий непосредственного использования языка запросов. Трудности возникают на уровне понимания схем данных и использования логических операторов. В частности, преподавательский опыт одного из авторов показывает, что даже студенты старших курсов, специализирующиеся в области информатики, при выполнении задания типа «*сделать запрос, выдающий данные за 3 и 5 октября*», нередко связывают даты логическим оператором «И».

Развитыми возможностями построения поисковых предписаний обладают, как правило, специализированные ИПС, справочно-информационный фонд которых содержит хорошо структурированные поисковые образы документов, причем возможности поискового интерфейса напрямую зависят от априорно оцениваемой возможности построения рядовыми пользователями сложных логических запросов. Так, в уже упоминавшейся базе данных журнала «Zentralblatt MATH», предназначено для профессиональных математиков, функция «Расширенный поиск» позволяет соединять в поисковом предписании при помощи логических связок до 5 значений элементов метаданных (притом сами эти элементы, с возможными их повторениями выбираются пользователем самостоятельно из общего списка), дополнительно указывая тип искомого документа и временной интервал его публикации.

И все же нельзя не отметить, что умение формально записать поисковый запрос, пусть и весьма сложный, – дело, собственно говоря, не слишком-то хитрое, требующее лишь известного опыта и небольших технических навыков. Гораздо нетривиальнее задача правильно выразить свою информационную потребность, т. е. *неформально* задать «характеристики предметной области, значения которых необходимо установить для выполнения поставленной задачи в практической деятельности» (ГОСТ 7.73-96).

Наиболее простая ситуация возникает, когда пользователь хочет найти конкретный документ, адрес хранения которого, однако, неизвестен. В этом случае задание в поисковом предписании в качестве ключевых слов имени автора документа и его названия, как правило, позволяют довольно быстро добиться нужного результата, даже если ИПС не предоставляет возможность структурировать вхождение перечисленных ключевых слов применительно к соответствующим полям метаданных. В последнем случае наибольшие проблемы могут возникнуть, если искомый документ относится к разряду «хрестоматийных» (как, например, «Гамлет» У. Шекспира, «Фауст» И. В. Гете или «Евгений Онегин» А. С. Пушкина) и существует масса документов, просто *упоминающих* о нем. Один из эффективных приемов решения подобной проблемы состоит в дополнении поискового предписания какой-либо достаточно длинной *цитатой* из текста (по возможности, не самой общеупотребительной).

Однако на практике пользователю обычно требуется найти не какой-то конкретный, заранее известный документ, а некие *сведения* (*факты*), знание которых необходимо для решения поставленной задачи (или же для удовлетворения любопытства). Возникающая при этом ситуация напоминает сюжет известной русской сказки «Пойди туда – не знаю куда, принеси то – не знаю что» (впрочем, подобные сказки известны в фольклоре многих народов мира – от Ирландии до Китая [Народные русские сказки..., 1985]), причем акцент ставится на первой части фразы, поскольку о том, что́ именно ему нужно, пользователь все-таки имеет некоторое представление. Сказочного Федота-стрельца вел к цели волшебный мячик. А как же следует составить поисковый запрос, чтобы скорее достигнуть поставленной цели?

«Лобовая атака» в форме постановки прямого запроса типа «Какова девичья фамилия жены М. Е. Салтыкова-Щедрина?» обычно не приведет к желаемому результату, поскольку современный уровень развития поисковых систем общего назначения не предполагает диалога с пользователем на естественном языке. Отметим, что поставленный выше вопрос – не совсем тривиальный, ибо ответы на «совсем тривиальные» вопросы типа «Где родился М. Е. Салтыков-Щедрин?» поисковые системы обычно все-таки находят, поскольку подавляющее большинство биографий писателя начинаются примерно так: «М. Е. Салтыков-Щедрин родился в январе 1826 года в селе Спас-Угол Тверской губернии» (слово «где» как служебное поисковой системой во внимание обычно не принимается). Кроме того, создатели некоторых веб-страниц, содержащих часто разыскиваемую в Сети информацию (обычного

не научного, а «бытового» характера), иногда включают предполагаемый вид пользовательского запроса (точнее, вопроса) в поисковый образ документа.

Более надежным способом составления поискового предписания представляется включение в поисковый образ запроса ключевых слов (или словосочетаний), которые, по мнению пользователя, непременно должны входить в текст документа, содержащего нужные сведения. Однако здесь возникает следующая дилемма: если включить в поисковый запрос небольшое количество «наиболее вероятных» слов, то его результатом будут сотни (а то и тысячи) документов, далеко не все из которых будут содержать ответ именно на поставленный вопрос. Если же включить в запрос много «предполагаемых» ключевых слов (или даже целую фразу), то мы рискуем получить на выходе пустое множество документов, поскольку авторы документов требуемой тематики могли описывать интересующий пользователя предмет фразами, несколько отличающимися от заданной в запросе.

Итак, в процессе поиска документов, содержащих некие интересующие нас факты, стоит задача сформулировать поисковое предписание таким образом, чтобы получить в результате его выполнения непустое множество документов, в котором процент «нужных» документов как можно более велик. Это резко повышает шансы сократить количество документов, просмотренных «впустую», т. е. прежде чем мы наткнемся на «нужный» документ. Проблемы, связанные с получением количественных оценок эффективности поиска, будут рассмотрены ниже.

О поиске «по аналогии»

В предыдущем пункте мы рассматривали ситуацию, когда поисковый образ запроса задается пользователем как некое «идеальное представление» о поисковом образе искомого документа. Однако, как уже отмечалось в начале статьи, информационные потребности научных работников, когда они в процессе исследования находятся на этапах изучения уже имеющихся в данной области результатов и научного поиска, характеризуются невысокой четкостью осознания и выражения. Опять-таки имеет место ситуация «Пойди туда – не знаю куда, принеси то – не знаю что», однако теперь уже акцент ставится на второй части фразы, поскольку известно, что описания документов, относящихся к той или иной научной тематике, заносятся в соответствующие реферативные базы данных. С другой стороны, у каждого исследователя за годы его работы образуется картотека библиографических описаний статей, книг и т. д., представляющих для него интерес. Основной критерий их отбора – личные интересы ученого. В настоящее время такие картотеки хранятся, как правило, на электронных носителях.

Таким образом, возникает задача нахождения по данному множеству документов класса схожих по содержанию документов (поиск «по аналогии»). В качестве информационного запроса предполагается задание непустого множества документов, а в качестве результата выполнения запроса выдаются документы, каждый из которых в определенном смысле близок к одному из документов, входящих в заданное множество. Процесс разбиения множества документов электронной базы на классы, при котором элементы, объединяемые в один класс, имеют большее сходство, нежели элементы, принадлежащие разным классам, называется *кластеризацией*.

Количественная характеристика меры сходства определяется на множестве документов D следующим образом:

$$m : D \times D \rightarrow [0,1],$$

причем функция m в случае полного сходства принимает значение 1, в случае полного различия – 0. Вычисление меры сходства осуществляется по формуле вида

$$m(d_1,d_2) = \sum a_i \, m_i \, (d_1,d_2), \qquad (1)$$

где i – номер элемента (атрибута) метаданных документа, a_i – весовые коэффициенты, причем $\sum a_i = 1$, $m_i(d_1,d_2)$ – мера сходства по i-му элементу (иными словами, по i-й шкале). Поскольку в описываемой ситуации практически все шкалы – номинальные (состоящие из дискретных текстовых значений), то мера сходства по i-й шкале определяется следующим образом: если значения i-ых атрибутов документов совпадают, то мера близости равна 1,

иначе 0. При этом необходимо учитывать, что значения атрибутов могут быть составными. В таком случае $m_i = n_{i1}/n_{i0}$, где $n_{i0} = \max\{n_{i0}(d_1), n_{i0}(d_2)\}$, а $n_{i0}(d_j)$ – общее количество элементов, составляющих значение i-го атрибута документа d_j, n_{i1} – количество совпадающих элементов. Заметим, что в качестве шкал целесообразно использовать следующие элементы метаданных: авторы, ключевые слова, текст аннотации. Кроме того, при задании меры можно принять во внимание тот факт, что значения весовых коэффициентов в формуле (1) определяются предполагаемой апостериорной достоверностью данных соответствующей шкалы и в определенных случаях один из коэффициентов может быть увеличен с пропорциональным уменьшением остальных. Например, полное (или даже «почти полное») совпадение значений атрибута «авторы» документа d_1 и документа d_2 более весомо в случае, когда количество значений этого атрибута в документе d_1 достаточно велико (по сравнению со случаем, когда документ d_1 имеет всего одного автора).

Основная проблема кластеризации документов заключается в таком разнесении документов по группам, при котором элементы каждой группы были бы настолько сходны друг с другом, чтобы в некоторых случаях можно было пренебречь их индивидуальными особенностями. При кластеризации документов важно прийти к разумному компромиссу относительно размера кластеров, избегая как формирования большого числа очень мелких кластеров (что снижает эффективность кластеризации как выделения множеств схожих документов), так и небольшого количества очень крупных классов (что может вызвать уменьшение точности поиска). Исследование различных алгоритмов кластеризации документов с целью выявления оптимального алгоритма для разбиения массива записей электронной базы с информацией о научных публикациях, на кластеры, содержащие в себе статьи по сходной тематике, проведено в работе [Барахнин и др., 2008].

Оценка эффективности поиска

Два основных понятия, в которых дается оценка эффективности поиска, определены в ГОСТ 7.73-96, причем эти определения остались практически неизменными с 1960-х гг. [Михайлов и др., 1968. С. 282–283]: *релевантными* называются документы, содержание которых соответствует информационному запросу, а *пертинентными* – содержание которых соответствует информационной потребности.

Разумеется, два этих понятия хотя и близки, но отнюдь не эквивалентны. Источник появления в выдаче нерелевантных документов – ошибки в описаниях и программном коде поисковых систем, а также прочие организационно-технические причины. При этом в тех случаях, когда поиск производится путем задания конкретного поискового запроса, возможно объективно судить о релевантности того или иного документа, вошедшего в выдачу, поскольку причиной выдачи нерелевантных документов (совокупность которого называется поисковым шумом) являются погрешности в индексировании документов (ручном или автоматическом), проявляющиеся, например, во внесении в поисковый образ документа «лишних» слов. Такая ситуация может возникнуть не только в результате явных ошибок, но и «языковых коллизий». Например, слова «вино» и «вина» имеют в некоторых падежах совпадающие словоформы, вследствие чего в поисковый образ документа, содержащего выражение «в вине», при автоматическом индексировании (которое, как правило, не сопровождается семантическим анализом текста) будут включены оба названных слова. Тем самым при включении в поисковый запрос слова «вино» будут выданы, в том числе, документы, содержащие слово с начальной формой «вина», которые являются, вообще говоря, нерелевантными. Обратите внимание, что при построении примера мы не могли ограничиться простыми омонимами, поскольку, например, при запросе «лук» релевантными будут документы как об оружии, так и о растении.

В тех же случаях, когда поиск производится «по аналогии», оценка релевантности документа носит более субъективный характер, поскольку такой поиск допускает произвол в способе задания меры сходства, в установлении ее порогового значения, отделяющего «похожие» документы от «непохожих» и т. п. Но даже если мы сочтем все эти параметры неотъемлемой частью поискового предписания, т. е. декларируем их «объективный» (для данного конкретного предписания) характер, то все равно останется практически неустрани-

мая зависимость результата поиска «по аналогии» от всей совокупности документов, входящих в информационный массив. Попросту говоря, вывод о схожести объекта «кошка» с объектом «корова» различается в случае, когда «информационный массив» есть множество лев, корова, и в случае, когда «информационный массив» – корова, кобра (или даже лев, корова, кобра).

Что же касается пертинентности, то понятие это – сугубо субъективное, поскольку потребности (не обязательно информационные) разных людей, пусть даже и выраженные одними и теми же словами-запросами, могут быть весьма различны. Так, потребность в *супе* с точки зрения среднестатистического русского удовлетворяется посредством щей или борща, а с точки зрения среднестатистического француза – посредством супа-пюре.

Уже из этого примера видно, что пертинентность выдачи может быть повышена посредством коррекции поискового предписания, формулируемого в соответствии с предполагаемым пониманием соответствующей потребности информационной системой (или, если угодно, разработчиками системы). Яркой иллюстрацией этого тезиса служит известный анекдот, в котором на вопрос пролетавших над незнакомой местностью воздухоплавателей: «*Где мы находимся?*» прохожий-математик дал абсолютно релевантный, но не пертинентный ответ: «*В корзине воздушного шара*». Конечно, объектом шутки здесь является буквализм математика, но ведь именно такое поведение характерно и для компьютерных алгоритмов. Поэтому правильно сформулированый запрос типа: «*Каковы наши географические координаты?*» или (если уж ориентироваться как на буквалиста, так и на обычного прохожего): «*Вблизи какого населенного пункта мы пролетаем?*» мог бы привести к пертинентному ответу.

В заключение перечислим основные количественные характеристики информационного поиска:

- коэффициент полноты: отношение числа найденных релевантных документов к общему числу релевантных документов, имеющихся в информационном массиве:

$$Recall = \quad |D_{\text{rel}} \cap D_{\text{retr}}| \,/\, |D_{\text{rel}}|,$$

где D_{rel} – множество релевантных документов в информационном массиве, а D_{retr} – множество найденных документов;

- коэффициент точности: отношение числа найденных релевантных документов к общему числу документов в выдаче:

$$Precision = \quad |D_{\text{rel}} \cap D_{\text{retr}}| \,/\, |D_{\text{retr}}|;$$

- коэффициент шума: отношение числа нерелевантных документов в выдаче к общему числу документов в выдаче:

$$Noise = |D_{\text{nrel}} \cap D_{\text{retr}}| \,/\, |D_{\text{retr}}|,$$

где D_{nrel} – множество нерелевантных документов в информационном массиве.

Заметим, что ни точность, ни полнота, взятые отдельно, не гарантируют высокого качества поиска. Так, выдача всех документов, имеющихся в информационном массиве, даст значение коэффициента полноты, равное 1, но точность при этом будет невысокой. Напротив, если выдан только один документ, и притом релевантный, то коэффициент точности равен 1, но при большом количестве ненайденных релевантных документов коэффициент полноты будет очень мал. Чтобы соблюсти баланс между полнотой и точностью, на практике используют так называемую F-меру (меру Ван Ризбергена), являющуюся средним гармоническим полноты и точности:

$$F = 2 \times Recall \times Precision \,/\, (Recall + Precision).$$

Заключение

Итак, мы проделали краткий экскурс в вопросы истории автоматизации информационного поиска, ознакомились с основными принципами работы современных информационно-поисковых систем и приемами построения поисковых предписаний и, наконец, изложили основные подходы к оценке эффективности поиска. Нетрудно заметить, что современное развитие алгоритмов информационного поиска характеризуется усложнением и даже «интеллектуализацией» поисковых алгоритмов. Вероятнее всего, в будущем ключевым термином станет раскопка текстов (text mining), иногда называемая аналитикой текстов (text

analytics) или раскопкой контента (content mining). А значит, в перспективе мы станем свидетелями конвергенции науки об информации и компьютерной науки.

Список литературы

Арский Ю. М., Гиляревский Р. С., Туров И. С. и др. Инфосфера: информационные структуры, системы и процессы в науке и обществе. М.: ВИНИТИ, 1996.

Барахнин В. Б., Нехаева В. А., Федотов А. М. О задании меры сходства для кластеризации текстовых документов // Вестн. Новосиб. гос. ун-та. Серия: Информационные технологии. 2008. Т. 6. вып. 1. С. 3–9.

Барахнин В. Б., Федотов А. М. Ресурсы сети Интернет как объект научного исследования // Изв. вузов. Проблемы полиграфии и издательского дела. 2008. № 1. С. 70–77.

Ляпунов А. А. Проблемы теоретической и прикладной кибернетики. Новосибирск: Наука, 1980. С. 320–323.

Михайлов А. И., Черный А. И., Гиляревский Р. С. Основы информатики. М.: Наука, 1968.

Народные русские сказки под редакцией А. Н. Афанасьева в трех томах. М.: Наука, 1985. Т. 2

Отле П. Библиотека, библиография, документация: Избранные труды пионера информатики / Пер. с англ. и фр. М.: ФАИР-ПРЕСС, Пашков дом, 2004.

Федотов А. М. Парадоксы информационных технологий // Вестн. Новосиб. гос. ун-та. Серия: Информационные технологии. 2008. Т. 6, вып. 2. С. 3–14.

Черняк Л. Статистическая машина Эмануэля Гольдберга // Открытые системы. 2004. № 03.

Шрейдер Ю. А. Равенство, сходство, порядок. М.: Наука, 1971.

Bush V. As We May Think. The Atlantic Monthly, 1945.

Otlet P. Traité de documentation. Bruxelles: Mundaneum, 1934.

Price D. J. de Solla. Little Science, Big Science. N.Y.; L.: Columbia Univ. Press, 1963 / Рус. пер. Д. Прайс Малая наука, Большая наука // Наука о науке. М.: Прогресс, 1966. С. 281–385.

Salton G. Dynamic Information and Library Processing. N. J.: Prentice Hall, 1975 / (Рус. пер. Дж. Солтон. Динамические библиотечно-информационные системы. М.: Мир, 1979).

Task Force on Metadata. Summary Report // American Library Association. 1999. June.

Материал поступил в редколлегию 26.05.2009

A. M. Fedotov, V. B. Barakhnin

PROBLEMS OF INFORMATION RETRIEVAL: HISTORY AND TECHNOLOGIES

Problems of information retrieval in Internet: history, technologies and algorithms are resolved in the article.

Keywords: information retrieval, internet.