УДК 681.3

Ю.В. Рогушина

ИСПОЛЬЗОВАНИЕ ОНТОЛОГИЧЕСКОГО ОПИСАНИЯ ПРЕДМЕТНОЙ ОБЛАСТИ ДЛЯ ПОВЫШЕНИЯ РЕЛЕВАНТНОСТИ ИНФОРМАЦИОННОГО ПОИСКА

Проведен анализ средств представления информации, публикуемой в Интернете. Предложен онтологический подход к описанию предметной области, интересующей пользователя, с целью задания контекста информационного запроса для повышения релевантности его результатов. Предлагаются различные способы интерпретации отношений между терминами онтологии, которые входят в контекст поиска.

Введение

Деятельность отдельных людей и организаций сейчас все больше зависит от имеющейся у них информации и способности ее эффективно использовать. Одним из основных средств получения информации сегодня является глобальная сеть Интернет — динамичная гетерогенная распределенная среда. Эффективный поиск информации в Интернете по мере увеличения объема и рассредоточения ее источников становится все более сложным и трудоемким. При этом критичным является не столько время поиска, сколько отбор информации, релевантной запросу пользователя.

Запрос пользователя представляет собой описание информации, доступ к которой он хочет получить. Такой запрос может, например, содержать ключевые слова, связанные логическими операторами; документ-образец; тип документа и его тему по классификатору; списки рекомендованных или запрещенных пользователем информационных источников; ограничения на время или объем поиска; объем, время создания, язык искомого документа. *Релевантность* результатов поиска оценивается с точки зрения пользователя. Документы, которые по какому-либо параметру, в том числе и не указанному явно в запросе, не удовлетворили пользователя, считаются нерелевантными. Чем сложнее форма представления запроса, тем более релевантные результаты можно получить в результате его выполнения. Однако усложнение формы запроса приводит к усложнению процедуры его обработки и, следовательно, к увеличению времени поиска.

Информационный поиск представляет собой процесс сопоставления запроса пользователя со сведениями об информационных ресурсах (ИР), известных информационно-поисковой системе (ИПС), к которой поступил этот запрос. В настоящее время доступ к информации, размещенной в глобальной сети Интернета, в подавляющем большинстве случаев обеспечивается при помощи поисковых машин — по оценкам, к поисковым службам обращается 71% всех пользователей Интернет [1]. Для формирования базы данных (БД) об ИР ИПС либо самостоятельно их индексирует, либо использует БД других ИПС. Второй способ значительно менее трудоемок, однако то, что структура этих БД уже задана разработчиками, ограничивает параметры поиска.

Постановка задачи. Эффективность выполнения поиска оценивается с точки зрения пользователя и определяется как соотношение между количеством документов, по его мнению, релевантных запросу, и общим количеством предложенных ему в результате выполнения поиска. Она зависит как от средств представления запроса и знаний об ИР, так и от способов их сопоставления и представления информации о конкретном пользователе, которая явным образом не указывается в запросе.

В данной статье рассмотрены средства представления информации в

Интернете и способы их индексирования, а также механизмы выполнения информационных запросов. На основе проведенного анализа предложены средства повышения эффективности поиска информации с помощью формирования *контекста запроса* — информации о пользователе, сфере его информационных интересов, предыстории ранее выполненных им запросов, различных предпочтениях и т.д. (например, пользователя может не заинтересовать документ, который он уже получал ранее в ответ на другой запрос, он не хочет пользоваться информацией из платных источников или не может обрабатывать информацию в определенных форматах). Для формализованного описания предметной области (ПрО), интересующей пользователя, предлагается использовать онтологические системы. Кроме того, в данной статье предлагается, чтобы пользователь не только применял готовые онтологии, но и имел средства для самостоятельного создания и модификации онтологии, которая отражает именно его представления о предметной области поиска. Это требует несколько упрощенного представления онтологических систем, но позволяет значительно расширить круг потенциальных пользователей, не нуждающихся в специальных знаниях в области информационных технологий и математической логики.

1. Информационные ресурсы Интернета

Для того чтобы найти адекватные средства информационного поиска, следует четко определить, среди каких именно объектов производится этот поиск и какие параметры этих объектов можно использовать для их идентификации. Ниже рассматриваются формы представления как текстовых, так и мультимедийных данных в Интернете, а также способы их индексации в БД различных ИПС.

1.1. Средства представления информации на естественных языках. Понятие гипертекста было введено В. Бушем еще в 1945 году, однако всплеск активности вокруг этой технологии произошел лишь тогда, когда с развитием Интернета возникла реальная необходимость в механизме объединения больших объемов информационных ресурсов, представленных в виде нелинейного текста. С использованием гипертекстовой модели документа представление разнообразных информационных ресурсов в сети стало более упорядочен, а пользователи получили удобный механизм поиска и просмотра нужной информации. Язык гипертекстовой разметки HTML, созданный специально для представления распределенной информации, является упрощенной версией стандартного описания формальных спецификаций разметки SGML [2] (Standard Generalized Markup Language — ISO 8879). Документ HTML состоит из стандартных элементов разметки, которые являются типизацией компонентов обычного документа — заглавие, авторы, параграфы, таблицы, цитирование и т.д. — и отображаются стандартным образом.

По мере увеличения количества информации в документах и усложнения их структуры простота технологии стала из достоинства превращаться в недостаток. Тенденцией современного развития Интернета является переход от документов, которые компьютер читает (machine readable), к документам, которые компьютер понимает (machine understandable), т.е. к обработке документов на семантическом уровне. XML (eXtensible Markup Language) [3] позволяет за счет расширения языка разметки явным образом выделить в документе структуру данных, что делает возможной дальнейшую машинную обработку документа, который при этом все еще остается понятным человеку, а также отделить данные, содержащиеся в документе, от того, каким образом документ будет представлен визуально. Технологии XML обеспечивают стандартное представление данных для обработки разными приложениями без специальной дополнительной обработки информации. Различ-

ные логические схемы разных документов могут использовать одни и те же имена элементов в различных значениях. Для интерпретации этих значений необходимо указать пространство имен — коллекцию имен, идентифицируемых по ссылке URI (URI — идентификатор ресурсов, позволяющий описывать и идентифицировать не только информационные ресурсы Интернета, но и предметы реального мира, общие понятия предметной области), которые используются документами XML в качестве имен типов, элементов и атрибутов. Пространство имен можно рассматривать как ИР, из которого извлекают необходимые определения.

Для описания ПрО, к которой относятся ИР, Консорциумом W3C в рамках Semantic Web — проекта семантической интерпретации ресурсов Интернета — предложен стандарт описания медатанных о документе RDF (Resource Description Framework) [4], который использует XML-синтаксис. Этот стандарт поддерживают многие ведущие производители программного обеспечения и поставщики контента. RDF описывает ресурсы в виде ориентированного размеченного графа — каждый ресурс может иметь свойства, которые в свою очередь также могут быть ресурсами или их коллекциями.

Однако для того, чтобы практически описать атрибуты документа, нужно дать им названия, которые потом будут использоваться во всём мире. В противном случае один автор напишет "Название", другой — "Заголовок", а третий — "Title“. В настоящее время наиболее распространен набор элементов для создания метаданных, разработанный международной группой Dublin Core Metadata Elements [5]. Он состоит из 15 элементов, которые можно условно разбить на три группы: *Content* — относящиеся к содержанию ресурса; *Intellectual Property* — характеризующие интеллектуальную собственность; *Instantiation* — описывающие конкретный экземпляр ресурса.

Этот набор элементов можно расширять, используя уже имеющиеся стандарты. Метаданные могут быть либо встроены в сам ИР, например в HTML-страницу (это самый простой подход для описания страниц), либо храниться и обновляться независимо от ИР. Второй подход более универсален, потому что в этом случае метаданные могут быть созданы для любого ресурса.

К сожалению, RDF-описания еще недостаточно широко распространены и для значительной части ИР отсутствуют.

Наряду с HTML часто применяются и другие форматы для представления текстовой информации. Например, PDF-файлы обычно не индексируются агентами ИПС. Между тем большой объем важной информации (в том числе технические статьи и научно-исследовательские отчеты) хранится только в формате PDF. Поэтому ведутся работы и в этом направлении. Так, система Google, дополненная новыми возможностями [6], может вести поиск примерно в 70% от общего количества PDF-файлов, опубликованных в Web. Google преобразует PDF-файлы в обычные текстовые документы, чтобы проиндексировать их как обычные Web-страницы.

В Интернете достаточно часто встречаются и материалы в форматах MS Word и rtf, в которых наряду с текстовой информацией содержатся рисунки, таблицы, графики и формулы. Преобразование в формат PDF не позволяет их дальнейшее редактирование, а в формат HTML — требует замены формул графическими изображениями, что делает их менее читабельными и также не позволяет их редактировать. Материалы в форматах MS Word и rtf практически не поддаются индексированию стандартными средствами ИПС. Последние версии MS Word предоставляют некоторые средства автоматизированного описания документов при помощи XML, но не все пользователи применяют их, а в большей части материалов, созданных

с помощью более ранних версий, такие описания отсутствуют.

1.2. Средства представления мультимедийных данных. Значительная часть ИР Интернета содержит наряду с текстовой информацией мультимедийные элементы: графику, видео, звук. Существует значительное количество широко распространенных форматов для хранения аудио- и видеоинформации, 3D-сценариев и изображений. Для того чтобы осуществлять поиск мультимедийных ИР, необходимо иметь адекватные средства как для их индексации, так и для описания искомого ИР. Это достаточно сложная задача, потому что графические и звуковые данные необходимо отразить в некое символьное представление, отражающее их семантику. Так, например, чтобы найти изображение людей на определенном фоне, нужно одно описание ИР, а чтобы найти изображение, на котором присутствуют математические символы, — совсем другое. Традиционные ИПС, которые развивались в тесной взаимосвязи с СУБД, в основном ориентированы на работу со структурированными текстовыми данными и мало приспособлены для обработки мультимедийной информации и данных, поступающих в оперативном режиме.

Альтернатива индексации естественноязыковой информации — технология, разработанная компанией Excalibur Technologies, которая объединяет метод адаптивного распознавания образов APRP (Adaptive Pattern Recognition Processing) и семантические сети. Она позволят работать с цифровой информацией любого типа — текстом, графикой, видео и др. Метод APRP опирается на теорию нейронных сетей и позволяет осуществлять бинарную индексацию, при которой размер индекса даже при обработке неструктурированной информации не превышает 30% от размера исходных данных [7].

Мультимедийные ресурсы значительно хуже, чем текстовая информация, поддаются индексации, т.к. использование методов, основанных на распознавании образов, требует очень больших вычислительных ресурсов. Поэтому достаточно часто (например, в ИПС Google и search.ua) для индексации изображений используются слова, содержащиеся в названии соответствующего файла, и текст подсказок. Но в ряде случаев такую индексацию нельзя считать удовлетворительной (в качестве названий иллюстраций часто используют обозначения типа «график 3» или «формула 5», не несущие практически никакой семантической нагрузки).

Многие современные ИПС предлагают услуги, относящиеся к категории «найти изображение, похожее на выбранное», но при этом качество их работы крайне низко (даже по сравнению с услугой «найти документ, похожий на выбранный», которая также работает недостаточно эффективно), а критерии отбора пользователю не ясны.

Если информация о мультимедийных ресурсах не представлена их поставщиками явным образом в какомлибо формате, известном средствам индексирования, то возникает необходимость в применении сложных и трудоемких операций (по распознаванию образов, речи и т.д.). Все возрастающий объем мультимедийной информации делает ее важным объектом для обработки средствами реферирования. Соответствующие технологии должны обрабатывать информацию из источников различных типов. Так, существующие методы работы с аудиоинформацией позволяют вычленить из потока информации законченные фрагменты (т.е. распознавать периоды тишины в разговоре, смену говорящего, снятие телефонной трубки и т.п.). Существуют также технологии обработки видеоинформации (определение ключевых элементов, логотипа), которые помогают определить тематику информации. Существуют системы, предназначенные для определения содержания видеофильмов путем распознавания шаблонов. Например, систе-

ма реферирования телевизионных новостей Broadcast News Navigator, опираясь на стратегию представления смешанной среды, объединяет ключевые кадры, автоматически извлеченные из видеофрагментов, и находит в них информацию об организациях, местоположении и участвующих в событиях лицах (наряду с такой информацией, как объем и время создания файла, длительность видеофрагмента и т.п.). Кроме того, для реферирования аудио- и видеоисточников информации широко применяются системы распознавания речи, после чего к сформированным естественноязыковым данным применяются средства автоматического реферирования текстовой информации.

В настоящее время группой MPEG (Moving Picture Experts Group [8]) разработан ряд стандартов для представление метаинформации о мультимедиа (например, MPEG7 [9] и MPEG21 [10]). MPEG-7 (Multimedia Content Description Interface — Интерфейс описания мультимедийных данных) обеспечивает стандартизацию описания разных типов мультимедиа для их поиска. Этот стандарт могут использовать как пользователи-люди, так и автоматические системы. Основной недостаток MPEG-7 — высокая сложность, поэтому для большей части мультимедийных ресурсов описание в этом формате отсутствует.

Несмотря на специфику мультимедийных ИР, наиболее приемлемым для осуществления информационного поиска (с учетом времени его выполнения и объемов хранимой в индексной БД информации) представляется их описание с помощью тех же средств, что и текстовой информации: ключевых слов, размера, даты создания файла и т.д.

1.3. Структурированные источники информации. При увеличении объема и усложнении структуры ИР возникает необходимость хранить информацию в БД, учитывающей особенности ПрО. При этом Web используется лишь как универсальный интерфейс пользователя с этой БД, а информация, предоставляемая конечному пользователю, формируется динамически (в ответ на действия пользователя соответствующие данные извлекаются из БД, а затем по ним формируется соответствующий документ).

Объем «глубинной» части Web (Deep Web) в 400—550 раз больше «поверхностной» (Surface Web) [11], и это соотношение продолжает увеличиваться, поскольку тенденция к хранению информации в структурированных источниках очевидна и по крайней мере в ближайшие годы не изменится. Локальный поиск по отдельному Web-серверу можно организовать несколькими способами. Если сервер меняется достаточно часто, то лучше использовать локальный поиск с помощью специализированной поисковой машины, которая устанавливается на Web-сервер и индексирует только его. Сейчас таких продуктов два: YandexSite компании CompTek и Следопыт компании MediaLingua. Еще одним способом организации локального поиска являются поисковые агенты, устанавливаемые на клиентскую машину и анализирующие информацию с Web-серверов. Они работают медленно, но позволяют более точно настроить механизм поиска и искать даже в тех местах, где поисковая машина не действует, например в корпоративной сети без выхода в Интернет. Хотя вся информация может быть найдена посетителем такого сайта при помощи локальной поисковой машины, глобальные поисковые машины, не приспособленные для работы с динамическим контентом, не способны проиндексировать информационные ресурсы сайта, вследствие чего потенциальный пользователь вообще не обратится к этому сайту.

2. Определение контекста поисковых запросов

Традиционные механизмы поиска в Интернете, как правило, рассматривают информационные запросы пользователя изолированно друг от друга и не учитывают полученные ра-

нее результаты. Имея информацию о пользователе, об интересующей его ПрО и о выполненных ранее запросах, можно получить более релевантные результаты и повысить эффективность поиска.

Существует несколько различных подходов к формализованному заданию таких сведений. Например, в проекте Inquirus [12] института NEC Research Institute контекстная информация задается явно в виде указания категории данных, которые запрашивает пользователь. Контекстная информация используется для выбора тех механизмов поиска, которым передается запрос, для модификации запросов и определения принципов упорядочения полученных документов.

2.1. Средства автоматического определения контекста поиска. Некоторые средства позволяют определить контекст поиска автоматически. Например, система Watson моделирует контекст на основе содержимого документов, которые пользователь ранее редактировал средствами Microsoft Word или просматривал в Internet Explorer. Эти документы анализируются с помощью эвристического алгоритма, который выявляет характерные слова, автоматически добавляемые к запросу. Кроме того, Watson в фоновом режиме ищет в Web документы, связанные с материалами, которые редактирует или просматривает пользователь. Недостатком системы является непрозрачность алгоритмов, используемых системой, для конечного пользователя.

Аналогично работает Remembrance Agent, который индексирует определенные файлы (сообщения электронной почты, научные статьи и т.п.) и, пока пользователь работает с некоторым документом, ведет поиск документов, связанных с ним. Autonomy's Kenjin [13] автоматически анализирует содержимое локальных файлов или файлов из Web, которые пользователь просматривает или редактирует. К аналогичным решениям можно отнести агентов Fab, Letizia [14] и WebWatcher, изучающих область интересов пользователя для того, чтобы предложить ему соответствующие Web-страницы.

2.2. Онтологический подход к представлению знаний о ПрО. Проблема информационного поиска усложняется тем, что различные сообщества людей используют в запросах специальные термины, имеющие различный смысл в разных ПрО (например, математическая *модель*, *модель* — уменьшенная копия технического устройства и фото*модель*). Так как большинство широко используемых ИПС являются не специализированными, а универсальными, то они не могут учитывать эти различия. В итоге значительная часть найденных ИР оказывается не релевантна запросу и пользователь должен сам просматривать большой объем не нужной ему информации. Специализированные же ИПС имеют довольно ограниченную информационную базу и, хоть и дают обычно высоко релевантные результаты поиска в определенной ПрО, не могут гарантировать обнаружение всех (или хотя бы значительной части) тех ИР, которые относятся к области их специализации и могут быть обнаружены универсальными ИПС. Таким образом, возникает противоречие между потенциальной доступностью публикуемой в Интернете информации и ограниченными возможностями человека по ее обнаружению.

Как показывает анализ публикаций, один из перспективных подходов к повышению эффективности поиска основывается на онтологиях (так, в проекте Semantic Web, направленном на анализ семантики ИР, именно онтологический подход [15, 16, 17] является основой для представления знаний о различных ПрО).

Понятие *онтологии*, заимствованное из философии, сейчас активно применяется в искусственном интеллекте и информационных технологиях. Основу онтологии составляют множество представленных в ней терминов и множество отношений между этими терминами [18]. Онтология — это некоторое описание взгляда на мир при-

менительно к конкретной области интересов, которое состоит из терминов и правил использования этих терминов, ограничивающих их значения в рамках конкретной ПрО. Использование онтологий способствует установлению корректных связей между элементами ПрО. Формальная модель онтологии O представляет собой упорядоченную тройку $O = \{X, \Re, \Phi\}$, где X — конечное множество концептов (понятий, терминов) предметной области, которую представляет онтология O; $\Re$ — конечное множество отношений между концептами заданной предметной области; Φ — конечное множество функций интерпретации, заданных на концептах и отношениях онтологии O [19].

2.3. Персонификация поиска при помощи онтологии ПрО, создаваемой конкретным пользователем. Повысить эффективность поиска позволяет его *персонификация*, т.е. использование сведений о предыдущих запросах конкретного пользователя и сфере его информационных интересов. Такой персонифицированный поисковый механизм может размещаться как на стороне сервера, так и на стороне клиента. Например, серверный механизм поиска Google способен отслеживать предыдущие запросы пользователя и выбранные им документы, а затем на основе этой информации делать вывод о сфере его интересов. Но из-за того, что затраты на работу такого механизма поиска очень высоки, полномасштабная персонификация на сервере сейчас обходится слишком дорого для основных механизмов поиска в Web. Большинство таких механизмов (исключение составляет лишь Northern Light [20]) даже не предлагают службу уведомления, которая сообщала бы пользователям о появлении новых страниц, соответствующих конкретным запросам [21, 22].

Наряду с глобальными онтологиями, которые описывают достаточно широкие ПрО и для создания которых необходимы значительные усилия как экспертов ПрО, так и инженеров по знаниям, существуют онтологии, позволяющие формально представлять знания конкретного пользователя о ПрО. Такие онтологии могут создаваться и модифицироваться пользователями самостоятельно. Хотя, возможно, некоторые представления пользователя о ПрО являются ошибочными, но такая онтология соответствует информационным интересам именно этого пользователя (например, если пользователь ошибочно считает дельфина рыбой и, запросив изображение какой-нибудь рыбы, получит изображение дельфина, то его информационная потребность будет удовлетворена).

Чтобы создать онтологию, пользователь должен задать конечное множество терминов ПрО, конечное множество отношений между этими терминами и конечное множество функций их интерпретации, а затем указать, между какими именно терминами существуют какие выражения (рис. 1). Онтология ПрО может быть визуализирована в виде леса ориентированных графов с нагруженными дугами, в котором вершины соответствуют терминам ПрО, а дуги — отношениям между ними.

3. Обработка результатов выполнения информационных запросов с учетом контекста

Для того чтобы пользователь имел возможность приступить к информационному поиску, ему надо предоставить непустое множество информационных ресурсов Q, $Q = \langle Q_1, ..., Q_n \rangle$, к которым он может обратиться. Такими ресурсами могут быть различные глобальные и локальные поисковые машины, отдельные сайты, фиксированные документы и т.д. Затем пользователь формирует информационный запрос. Способ выполнения поиска зависит от специфики конкретного ИР. В результате выполнения поиска формируется множество документов I, которые ИПС посчитали релевантными запросу.

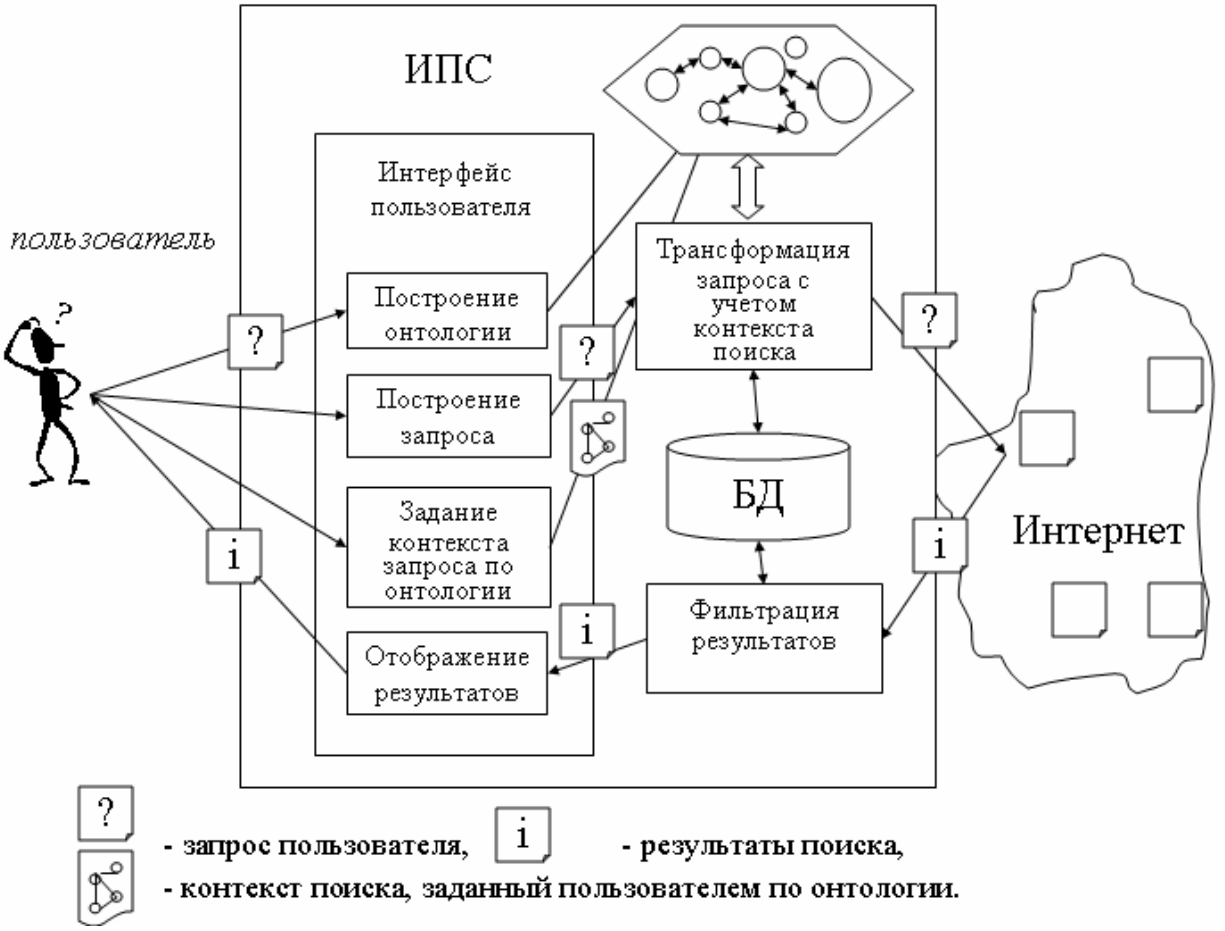

Рис. 1. Схема организации поиска на основе онтологии

$I = \bigcup_{i=1}^{n} I_j$, где I_j — результат поиска в информационном ресурсе Q_j.

К сожалению, большинство ИПС, осуществляющих поиск по ключевым словам, включают в I очень много ненужной информации — повторы, нерелевантные и устаревшие ссылки, а также ссылки на документы, уже известные пользователю. Чтобы избавить пользователя от необходимости просматривать вручную все эти документы, предлагается осуществить их фильтрацию, используя сведения о предыдущих запросах этого пользователя и сфере его информационных интересов.

3.1. Этапы обработки результатов выполнения запросов. Обработка результатов выполнения запросов состоит из 6 этапов (рис. 2).

Этап 1. В результате выполнения информационного запроса пользователя к Q по ключевым словам формируется множество I. Если доступна мета-информация о соответствующем ИР (например, в формате RDF или MPEG7), то поиск осуществляется с учетом этой информации.

Этап 2. Если множество I не пусто, выполняется упорядочение этого множества по URL-адресам ссылок. Иначе — завершение работы.

Этап 3. Если полученное на этапе 2 множество I_1 не пусто, отфильтровываются ссылки-"зеркала". Повторяющиеся адреса отбрасываются. Иначе — завершение работы.

Этап 4. Отфильтровываются устаревшие ссылки.

Этап 5. Если полученное на этапе 3 множество I_2 не пусто, осуществляется проверка по БД пользователя, получал ли он ранее каждую из оставшихся ссылок (если получал, то решение о том, оставлять ли эту ссылку, зависит от того, как в прошлом пользователь

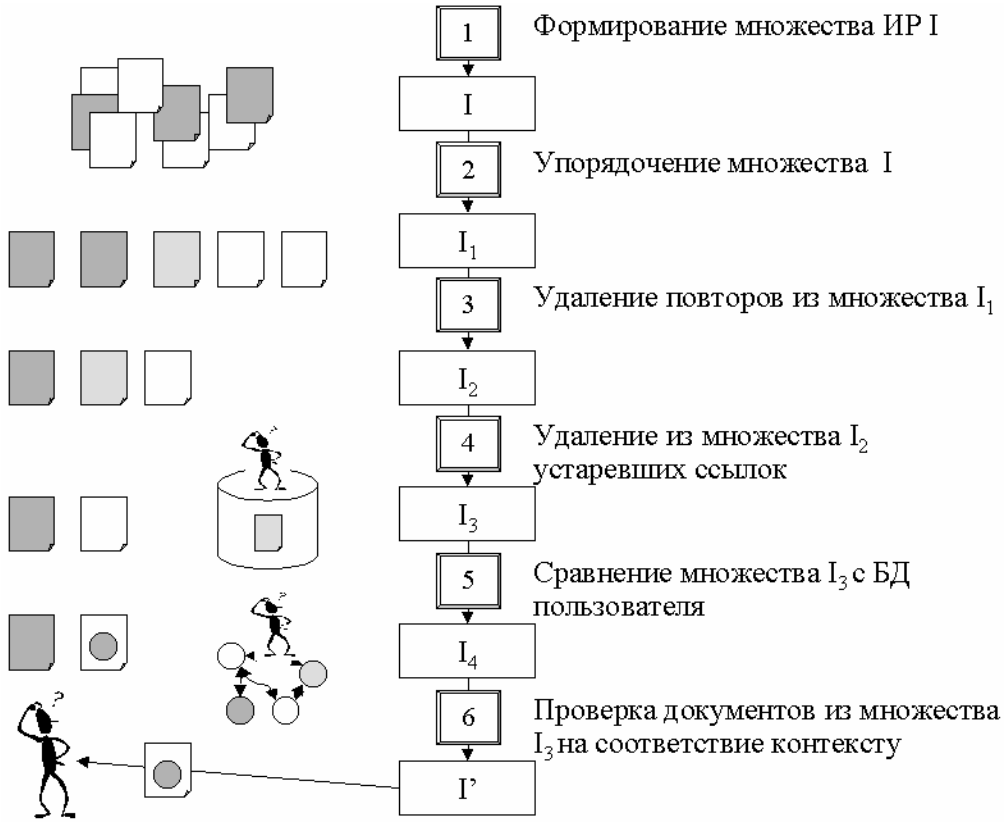

Рис. 2. Этапы обработки результатов выполнения запросов

поступил с этой ссылкой, а также от других его инструкций). Иначе — завершение работы.

Этап 6. Если сформированное на этапе 5 множество I_4, $I_4 \subseteq I$ не пусто, выполняется оценка соответствия документов i_j, $j = \overline{0,k}$ из этого множества контексту поиска. Иначе — завершение работы.

Именно на 6-м этапе используется онтология ПрО, созданная ранее пользователем. Здесь контекст поиска — это непустое неупорядоченное множество терминов и словосочетаний, характерных, по мнению пользователя, для того ИР, который он хочет найти. Так, например, наличие в ИР терминов «монография», «список литературы» и «аннотация» повышают вероятность того, что рассматриваемый ИР — научная работа.

Применение пользовательских онтологий для задания контекста поиска в первую очередь ориентировано на пользователей, имеющих постоянные информационные интересы в сети и требующих постоянного поступления соответствующей информации. Запросы таких пользователей могут повторяться от сеанса к сеансу или изменяться, но ПрО, в которой пользователи являются экспертами, практически не изменяются и являются достаточно ограниченными. Описание этих ПрО задается самими пользователями в виде онтологий. Один пользователь может создавать несколько онтологий, если он имеет несколько интересующих его прикладных областей, которые не пересекаются.

3.2. Использование онтологии, созданной пользователем, для предоставления контекста поиска. Онтология используется для предоставления контекста поиска — информации о ПрО, которая интересует пользователя, предыстории его запросов и другой информации о конкретном пользователе, его информационных предпочтениях. Это осуществляется следующим образом.

В онтологии пользователь может отметить термины, наличие которых в искомом документе является желательным или нежелательным, а также

задать более сложные операции (например, автоматически отметить все термины, находящиеся в заданном отношении с терминами, отмеченными ранее). Это позволяет, в частности, легко учитывать при поиске синонимы или близкие по значению слова, а также осуществлять поиск сразу на нескольких языках.

В результате формируется непустое множество слов (или словосочетаний) $W = \{w_1,...,w_m\}$, каждое из которых может иметь свой положительный либо отрицательный вес v_k, $k = \overline{1,m}$. Затем для каждого документа i_j, $j = \overline{0,k}$, из множества I', $I' \subseteq I$ формируется коэффициент соответствия контексту поиска

$$s_j, j = \overline{0,k}, s_j = \sum_{k=1}^{m} v_k * f(i_j, w_k), \qquad (1)$$

где $f(i_j, w_k) = \begin{cases} 1, \text{если } w_k \in i_j, \\ 0, \text{если } w_k \notin i_j. \end{cases}$

Чем выше коэффициент (1), тем, вероятно, выше релевантность документа запросу пользователя.

В некоторых случаях может быть полезно использовать более сложную формулу расчета коэффициента соответствия контексту поиска:

$$s'_j, j = \overline{0,k}, s'_j = \sum_{k=1}^{m} v_k * f(i_j, w_k) * t_k , \qquad (2)$$

где t_k, $k = \overline{1,m}$, — количество вхождений термина w_k, $k = \overline{1,m}$, в документ i_j, $j = \overline{0,k}$.

После выполнения оценки найденных ИР с помощью (1) или (2) пользователю в первую очередь предлагаются ИР, имеющие наиболее высокий коэффициент соответствия контексту поиска (фиксированное количество ИР или все найденные ИР, имеющие коэффициент соответствия контексту поиска выше определенной пользователем константы).

Пользователь может обращаться к онтологиям, созданным другими пользователями, — просматривать их, задавать по ним контекст поиска, копировать из них нужны фрагменты, но не имеет права изменять их. ИПС должна предусматривать поиск онтологий, которые содержат введенные пользователем термины, а также поиск онтологий, похожих на избранную пользователем онтологию. Это позволяет создавать группы пользователей с общими информационными интересами и предотвращать дублирование в выполнении одинаковых многоразовых запросов различных пользователей.

Для реализации информационного поиска в Интернете представляется целесообразным использование интеллектуальных программных агентов [23], позволяющих обращаться за информацией к локальным поисковым машинам сайтов без непосредственного участия пользователя, и разработка мультиагентной информационно-поисковой системы [24], в состав которой входят агенты информационных ресурсов, обеспечивающие интерфейс с локальными поисковыми системами различных сайтов, и агент-диспетчер, обеспечивающий перечень таких сайтов.

Заключение

Рассмотрев различные средства представления метаданных о разнородной (в том числе и мультимедийной) информации, которая публикуется в распределенной динамически изменяющейся среде Интернет, можно сделать вывод о том, что, несмотря на многообразие подходов к отражению семантики информационных ресурсов, на современном уровне развития информационных технологий в большинстве случаев наиболее релевантным и полным остается информационный поиск по ключевым словам.

Повышение эффективности такого поиска является сегодня актуальной задачей. Этого можно добиться путем обработки контекста запроса и сведений о конкретном пользователе, пославшем запрос, а также предыстории его обращения к различным ИПС.

Для формализованного описания предметной области поиска, к которой относятся информационные интересы

пользователя, целесообразно использовать онтологический подход. При этом необходимо создание адекватных платформонезависимых инструментальных средств для создания, модификации и обработки онтологических систем, которые может применять пользователь, не являющийся специалистом в области информационных технологий.

1. *Greenberg I., Garber L.* Searching for new search technologies // IEEE Comp. — 1999. — Aug. — P. 6–11.
2. *ISO 8879.* — http://www.iso.ch/cate/d16387.html.
3. *Extensible* Markup Language (XML) 1.0, W3C Recommendation. — http://www.w3.org/TR/-1998/REC-xml-19980210.
4. *RDF* Model Theory / W3C Working Draft. 2002. — http://www.w3.org/TR/rdf-mt/.
5. *RFC2413*: Dublin Core Metadata for Resoerce Discovery. — http://www.faqs.org/rfcs/-rfc2413.html.
6. *Найти* можно все // PC Magazine. — http://www._PCMagazine.CSS/Common.css.
7. *Картышева Е.* Интеллектуальные поисковые системы Excalibur. — http://www.osp.ru/-nets/1997/06/98.html.
8. *MPEG*: Achievements and Current Work. — 2001. — http://www.cselt.it/mpeg/terms_-of_reference.htm.
9. *MPEG-7*: Overview. — 2002. — http://mpeg.-telecomitalialab.com/standards/mpeg-7/-mpeg-7.htm.
10. *MPEG-21*: Overview. — 2002. — http://mpeg.telecomitalialab.com/standards/-mpeg-21/mpeg-21.htm.
11. *Deep* Web. — http://www.completeplanet.com/tutorials/deepweb.
12. *Architecture* of a metasearch engine that supports user information needs / E. Glover, S. Lawrence, W. Birmingham, C.L. Giles // Eighth International Conf. on Information and Knowledge Management, CIKM 99. — Kansas City, Missouri, 1999. — P. 210–216.
13. *Autonomy's* Kenjin. — http://www.kenjin.com.
14. *Lieberman H.* An Agent That Assists Web Browsing. — http://lieber.media.mit.edu.
15. *A Model-Theoretic* Semantics for DAML+OIL. — http://www.w3.org/TR/daml+oil-model.
16. *W3C* Web Ontology. — http://www.w3.org/-2001/sw/WebOnt/.
17. *Requirements* for a Web Ontology Language, W3C Working Draft. — http://www.w3.org/-TR/webont-req/.
18. *Россеева О.И., Загорулько Ю.А.* Организация эффективного поиска на основе онтологий. — http://www.dialog-21.ru/archive_-article.asp.
19. *Гаврилова Т.А., Хорошевский В.Ф.* Базы знаний интеллектуальных систем. — Спб.: Питер, 2001. — 382 с.
20. *Northern* Light. — http://www.Northern-Light.com/help.htm.
21. *DriveWay*. — http://www.driveway.com.
22. *Xdrive*. — http://www.xdrive.com.
23. *Рогушина Ю.В.* Программные агенты: определения, таксономии и модели // УСиМ. — 2001. — N 5. — С. 39–45.
24. *Рогушина Ю.В.* Разработка средств интеллектуализации поиска информации в Интернете // Пробл. программирования. — 2002. — N 1–2. — С.379–385.

Получено 18.06.03

Об авторе

Рогушина Юлия Витальевна,
кандидат физико-математических наук,
старший научный сотрудник

Место работы автора
Институт программных систем НАН Украины,
просп. Академика Глушкова, 40,
Киев-187, 03680, Украина
Тел. (044) 268 4698