OSTIS-2014
(Open Semantic Technologies for Intelligent Systems)

UDK 681.3.

ЗНАНИЕ-ОРИЕНТИРОВАННЫЕ СРЕДСТВА ПОДДЕРЖКИ СЕМАНТИЧЕСКОГО ПОИСКА В WEB

Рогушина Ю.В.

Институт программных систем НАН Украины, Киев, Украина
ladamandraka2010@gmail.com

В статье рассматриваются основанные на семантике направления развития поиска в Web. Проанализированы средства представления знаний (онтологии и тезаурусы), методы их пополнения и сопоставления. Предложены способы применения персональной и коллаботативной информации при семантическом поиске.
Ключевые слова: семантический поиск, онтологическая модель, тезаурус, индуктивный вывод, рекомендующие системы, коллаборативный поиск.

Введение

Значительная часть современных приложений ориентированы на функционирование в открытой информационной среде, в частности, на извлечение из ресурсов Web знаний об интересующей пользователя предметной области (ПрО). Таким образом, проблема поиска в Web оказывается составной частью самых разных информационных систем, а общая тенденция экспоненциального увеличения объема доступной через Web информации обуславливает потребность в переходе при информационном поиске от обработки больших объемов данных к обработке знаний – более компактных, но имеющих значительно более сложную структуру.

В связи с этим возрастает актуальность как методов извлечения знаний, так и их повторного использования. Основой для интероперабельного представления знаний может послужить онтологический анализ, базирующийся на стандартах и программных средствах, разработанных в рамках проекта Semantic Web.

То, что онтологическое моделирование является адекватным средством для описания различных ПрО, является на сегодня общепризнанным фактом, а широкий выбор онтологий, доступных через Web, подтверждает популярность этого подхода среди различных групп разработчиков и пользователей Web-приложений.

При этом возникает ряд вопросов, связанных с формализацией семантики, в частности, построения формализованной семантической модели знаний, которая обеспечивает однозначную и автоматизируемую их интерпретацию. На сегодня уже разработан и используется ряд достаточно эффективных средств семантического моделирования, позволяющих адекватно отображать сведения о реальном мире, например, RDF Schema, однако их выразительные возможности требуют расширения. Такие возможности предоставляют онтологии. Онтология – это система, состоящая из множества понятий и множества утверждений об этих понятиях, на основе которых можно строить классы, объекты, отношения и теории [Бездушный, 2003]. Онтологическое позволяет интероперабельно описать предметную область.

Важным вопросом является выбор языка для представления онтологий, который основывается на том, какие именно логические формализмы лежат в их основе. При этом онтология рассматривается как логическая теория.

Необходимо, чтобы логический язык предоставлял достаточно выразительных возможностей для формализации знаний в описываемой ПрО. Кроме того, следует учитывать разрешимость и вычислительную сложность тех онтологий, которые могут быть созданы при помощи такого языка, так как от этого зависит сама возможность создания программных средств для обработки таких онтологий и необходимые для этого вычислительные ресурсы. В настоящее время для описания онтологий используются такие логические формализмы, как логика предикатов первого порядка (в языке Ontolingua), фреймовая логика (в языке F-Logic) и дескриптивная логика (в языках DAML+OIL, OWL DL). Разрешимость логического языка гарантирует получение ответа от

системы логического вывод, но время его получения зависит от вычислительной сложности языка.

Чем выше выразительность логического языка, тем точнее он позволяет описывать ПрО, но тем больше времени затрачивается на логический вывод. Различные дескриптивные логики обладают различной выразительной мощностью. Поэтому использование логических языков представления онтологий на основе дескриптивных логик позволяет обеспечить приемлемый компромисс между выразительными возможностями языка и необходимыми для обработки таких онтологий вычислительными ресурсами: можно в зависимости от задачи выбрать язык с достаточной выразительностью и минимальной вычислительной ресурсоемкостью. OWL DL базируется на дескриптивной логике SHIQ и обеспечивает достаточно точное прогнозирование сходимости и времени работы различных методов обработки знаний, представленных в виде онтологии.

Онтологический анализ в семантическом поиске

При традиционном информационном поиске производится сопоставление запроса пользователя, характеризующего его информационную потребность, со сведениями о контенте проиндексированных информационных ресурсов (ИР), и в ответ на запрос пользователя формируется группа информационных объектов, известных ИПС и по тем или иным параметрам сопоставленных с запросом, т.е. $I = \{i_j, j = \overline{1,n}\} = f(z, DB_{инс})$.

При этом субъектами обработки являются пары (запрос, ИР), а сопоставление производится унифицированно для всех пользователей. Такой поиск обычно основывается на обнаружении лексического соответствия ключевых слов и терминов, содержащихся в документе, на основе минимальных предварительных знаний, рассматривая текст как набор слов

Как правило, такой поиск применяют для обработки неструктурированной текстовой информации. Кроме того, часто учитывается расстояние между обнаруженными терминами. Для структурированных документов имеет значение месторасположение ключевых слов (например, в заголовке или в метаописании).

Анализ исследований в области информационного поиска показал, что дальнейшее усовершенствование средств описания и алгоритмов сопоставления запросов и ИР практически не улучшает ситуацию, т.к. для более эффективного поиска нужно использовать дополнительные знания – о ПрО, интересующей пользователя, о сообществах пользователей с подобными интересами и о качестве ИР.

При персонифицированном поиске субъектами обработки являются уже тройки (запрос, пользователь, ИР) $I_{pers} = f(z, u, DB_{инс})$. При этом сопоставление производится уже по-разному для различных пользователей в зависимости от их персональных предпочтений. Сведения о пользователе хранятся в БД поисковой системы и могут содержать его формальные характеристики (например, перечень естественных языков, которые знакомы пользователю), историю ранее выполненных запросов и знания об интересующих его предметных областях.

К сожалению, не всегда опыта конкретного пользователя достаточно для того, чтобы определить полезность тех или иных ресурсов (например, при обращении к новой ПрО или к новой группе ресурсов). В таком случае полезным может оказаться опыт других пользователей, производивших ранее поиск в той же области и имеющих сходные информационные потребности. При коллаборативном поиске целесообразно использовать методы, применяемые в рекомендующих системах. Если мы говорим о семантическом поиске, то предполагается, что информация в процессе сопоставления должна обрабатываться на семантическом уровне, с использованием знаний (о пользователе, ресурсах, предметной области и т.д.).

При простом знание-ориентированном поиске при обработке учитываются не только формальные сведения о запросе, пользователе и ИР, но и более сложно структурированные знания о них. Тогда при их сопоставлении необходимо будет оценивать степень подобия этих знаний. В частности, при онтологическом подходе к представлению знаний для каждого ИР может указываться онтология ПрО, характеризующая его контент, а для пользователя – онтология интересующей его ПрО, а при сопоставлениии ИР и запроса необходимо будет выполнить сопоставление этих двух онтологий.

Семантический поиск определяется как метод информационного поиска, в котором релевантность документа запросу определяется семантически (по близости смысла), а не синтаксически (по встречаемости ключевых слов в документе).

То, какие именно знания используются, как они представлены и как они обрабатываются, зависит как от специфики разрабатываемой ИПС, так и от концепции, выбранной ее разработчиками, но в общем случае $I_s = \{i_j, j = \overline{1,n}\} = f(z, DB_{инс}, KB_{инс})$.

При этом часто предоставляются возможности для нечеткого поиска (например, обрабатываются неправильно написанные ключевые слова), поиска с учетом контекста

Если же речь идет о семантическом поиске в Web, то следует учитывать, что при этом в Web могут быть размещены не только информационные объекты, среди которых осуществляется поиск, но и внешние базы знаний, используемые при поиске. Поэтому при создании таких систем следует

учитывать, что такие внешние БЗ могут менять контент, структуру и доступность независимо от разработчиков ИПС.

$$I_{web\_s} = \left\{ i_j, j = \overline{1,n} \right\} = f\left(z, DB_{ипс}, KB_{ипс}, \left\{ KB_{web\_k}, k = \overline{1.m} \right\} \right).$$

Следует учитывать, что сегодня многие ИПС (например, Google) стремятся накапливать и использовать опыт взаимодействия с конкретным пользователем. Но часто информационные потребности пользователя оказываются ограниченными во времени (например, накопив информацию для выбора нового телефона, пользователь покупает его и больше не нуждается в сведениях о телефонах, а ИПС продолжает предлагать их ему) либо вообще не связанными с ним (например, запрос выполняется по чьей-либо просьбе). Кроме того, часть своих информационных потребностей пользователь не хочет делать открытой информацией – к примеру, запросы, связанные с отдыхом или здоровьем не хочет смешивать с запросами по работе. Поэтому более целесообразно при выполнении запросов дать возможность пользователю включить его в один из своих профилей либо вообще не сохранять для дальнейшей обработки

Таким образом, можно выделить основные направления развития поиска: 1. от формального – к семантическому; 2. от унифицированному – к персонифицированному; 3. от индивидуальному – к коллаборативному; 4. от закрытого – к управляемому; 5. от монотонного – к тематическому (с учетом динамики и конечности информационных потребностей). Наиболее полно удовлетворить информационные потребности пользователя позволяет интегрированное использование всех этих возможностей, т.е. персонифицированный, Web-ориентированный коллаборативный поиск, основанный на знаниях. Как правило, такой поиск является надстройкой над уже существующими поисковыми системами (например, Google), которая позволяет переупорядочить результаты поиска.

Чтобы повысить эффективность поиска, целесообразно как можно более точно определить, к какой ПрО относится информационная потребность пользователя и какой именно тип информации необходим для ее удовлетворения. В первом случае пользователю нужно выбрать одну из существующих онтологий ПрО и при необходимости модифицировать ее в соответствии со спецификой его проблемы. Так как не только создание и модификация онтологий, но и анализ содержащихся в уже созданных онтологиях знаний представляет собой достаточно сложную задачу, то такой подход приемлем только в том случае, если в дальнейшем такая онтология будет использоваться для выполнения единичного запроса, а для достаточно большого числа связанных с этой ПрО запросов. Такая ситуация характерна, например, для научных исследований, когда пользователь старается обнаруживать новые публикации и разработки по интересующей его проблематике. Для таких ситуаций характерно как достаточно глубокое и структурированное понимание пользователем специфики ПрО (что и позволяет ему создавать онтологическую модель ПрО), так и отсутствие полностью удовлетворяющей его потребностям и общепринятой онтологической модели ПрО, так как научные исследования проводятся именно в тех областях, где не для всех вопросов уже найдены удовлетворительные решения и именно с целью нахождения таких решений.

Во втором случае пользователю необходимо воспользоваться какой-либо онтологией или таксономией информационных объектов, информация о которых содержится в ИР Web. Например, воспользовавшись организационной онтологией, пользователь может указать, что ему необходимы сведения об организациях с теми или иными свойствами, о людях, имеющих определенную квалификацию, либо о проектах, решающих какие-то проблемы.

Но и этих сведений недостаточно в том случае, если условиям запроса удовлетворяют много различных ИР – для пользователя достаточно трудоемко самому анализировать все предложения информационно-поисковой системы (ИПС). Возникает проблема ранжирования списка найденных релевантных ИР. Кроме уже широко используемых унифицированных подходов, кода такое ранжирование будет одинаковым для всех пользователей, целесообразно учитывать как персональный опыт конкретного пользователя, так и опыт того сообщества пользователей, которое пользователь признает авторитетным для себя именно в этом запросе. Следует отметить, что в большинстве случаев ИПС, использующие опыт тех или иных сообществ, либо обобщают информацию, поступающую от всей совокупности пользователей, либо самостоятельно вычленяют некоторую подгруппу пользователей (кластер), которую считают подобно тому пользователю, для которого выполняется запрос. При этом возникает три проблемы: 1) как правило, пользователю непонятно, по каким критериям формируется такая группа; 2) группа формируется одинаково для любых запросов пользователя; 3) пользователю не предоставляются средства для явного формирования такой группы.

Постановка задачи

Чтобы обеспечить для различных интеллектуальных приложений эффективный доступ к ресурса открытой информационной среде Web, необходимо разработать интегрированную формальную модель, обеспечивающую интероперабельное представление знаний о пользователях, ресурсах и специфике предметных областей (в частности, учитывающих Semantic Web и социальный Web). При разработке методов обработки представленных в этой модели знаний необходимо проанализировать уже существующие средства онтологического анализа, возможности

использования тезаурусов, методы индуктивного извлечения знаний и алгоритмы выработки рекомендаций на основе накопленного сообществом пользователей опыта об интересующей пользователя предметной области.

Семантический поиск в Google

В последнее время многие разработчики ИПС в той или иной степени декларируют применение онтологий и поддержку семантического поиска.

Рассмотрим, что понимают под семантическим поиском в Google его разработчики. Это достаточно важно, так как на сегодня именно Google является наиболе широко используемым средством поиска в Web. Семантический поиск Google включает три основных компонента [Amerland D., 2013]: адрес URI; RDF; значение (онтологию). При этом онтологии позволяют связывать друг с другом различные данные, описанные при помощи RDF и адресованные через URI. Кроме того, для поддержки семантического поиска в Google применяется новый инструмент, обеспечивающий семантический поиск, – Knowledge Graph (Сеть знаний), который обеспечивает связь между различными элементами проиндексированного контента ИР и позволяет объединять информацию из различных источников.

Knowledge Graph позволяет непосредственно на странице с результатами поиска получать информацию об объекте поиска и связанные с ним факты: справа от результатов поиска на экран выводится информационная панель, отображающая сведения о географических объектах, людях, фильмах и т.п. Это позволяет пользователю получить информацию, не переходя на сайт, послуживший источником информации.

Сейчас в базе находится (по различным оценкам) от 500 до 700 миллионов объектов и около 3.5 миллиардов связей между ними. Объем знаний, доступный различным национальным версиям инструмента, различается. Информация поступает в базу данных сервиса в основном из открытых источников: для России в основном из русскоязычной Википедии, а в США используется также онлайн-справочник CIA World Factbook и база структурированных данных Freebase, собранная из множества отдельных вики-проектов: Google в 2010 году приобрел компанию Metaweb Technologies – разработчика базы данных Freebase, которая на тот момент содержала около 12 млн. сущностей.

Предполагается, что Knowledge Graph позволяет поисковой системе понимать смысл запроса. При помощи Knowledge Graph Google не просто идентифицирует ключевые слова, но и понимает смысл поискового запроса и ищет информацию о нем в базе данных Google, которая содержит сведения о различных объектах — например, людях, местах, вещах, фильмах, произведениях искусства и прочем. Сеть знаний является первой ступенью на пути Google к интеллектуальному поиску.

В Web доступна информация, поступающая из Web-сайтов, социальных сетей, локальных сетей, профилей, баз данных и собственно от Google Search. Однако этой информации недостаточно. Google использует правила вывода, чтобы определить, как информация объединяется в группы и что она значит. Эти правила формируются на основе того, как пользователи работают с данными в социальных сетях (распространяемый контент, комментарии и взаимодействие), форумах и при поиске в Google. Кроме того, отслеживается поведение пользователей на различных Web-сайтах и анализируются последовательности их действий. Важным источником информации служат также сервисы, определяющие местонахождение пользователей по сигналу GPS от их мобильных устройств, определение географического местоположения по IP-адресу и сведения, получаемые при взаимодействии устройств.

Основная причина, по которой Google отображает на своей странице с результатами поиска прямые ответы на запросы – дольше удерживать на этой странице пользователя, т.к. это время является важным параметром для рекламодателей. Конкуренты Google, в первую очередь Facebook, заметно упрочили свои позиции в борьбе за внимание пользователей — а следовательно, и за деньги рекламодателей: посещение Facebook занимает около 13% того времени, которое средний американец проводит в Сети, а на долю сервисов Google, включая YouTube, остается лишь 11%. Помимо этого Facebook в значительной степени перенял у Google ее бизнес-модель: персонализированные предложения в этой социальной сети — это не что иное, как контекстная реклама Google, формирующаяся под влиянием запросов и интересов пользователей.

Google подчеркивает, что персонализация поиска в сочетании с семантическим поиском дадут новый вариант релевантности выдачи. Тем не менее, следует отметить, что сейчас такая информация для запросов – не только на русском и украинском языках, но и на английском – предоставляется в ответ только на самые короткие и простые запросы и по охвату значительно меньше, чем набор статей в Википедии. Поэтому остается открытым вопрос, насколько будет полезен пользователям новый сервис Google и насколько на самом деле в нем задействован семантический поиск. Главная проблема для пользователей – отсутствие явной модели поиска и сведений об онтологии, используемой при поиске, что позволялл бы хоть в какой-то мере прогнозировать его результаты.

Кроме того, непонятно, какую политику применяет Google в тех случаях, когда ответы не однозначны и зависят не только от местоположения пользователя, языка, на котором он говорит, времени года и т.п., но и от более сложных и субъективных факторов (например, запрос «самые

красивые собаки» или «выдающиеся политики», на которые не может быть однозначного ответа).

Онтологическая модель взаимодействия пользователей и ресурсов в Web

Семантический поиск представляет собой надстройку над традиционным информационным поиском, в котором с целью повышения пертинентности поиска используется обработка знаний, которые касаются как самого пользователя и его информационных потребностей (персонификация поиска), так и об информационных ресурсах, среди которых осуществляется поисковая процедура. Таким образом, семантический поиск состоит из формирования информационных моделей пользователя, интересующей его предметной области, задачи, которую решает пользователь, и информационных моделей доступных информационных ресурсов (ИР), которые характеризуют их семантику, и их дальнейшего сопоставления.

Сейчас для интероперабельного представления различных знаний в Web широко применяются онтологии, обеспечивающие явное формализованное представление семантики представленной информации и обеспечивающие возможность логического вывода на них. Поэтому представляется целесообразным для поддержки семантического поиска разрабатывать именно онтологические модели взаимодействия пользователей и ИР в информационном пространстве Web, а также методы их сопоставления и пополнения.

Онтологическая модель пользователя представляет собой класс онтологии, экземплярами которого являются сведения о зарегистрированных в ИПС пользователях (рис.1).

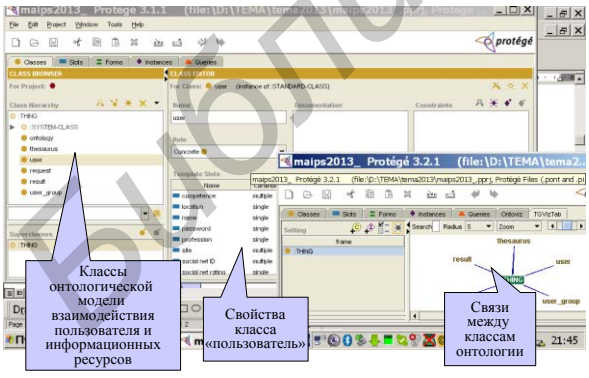

Рисунок 1 - Онтологическая модель взаимодействия пользователей и ИР

В онтологической модели взаимодействия пользователей и ИР описаны следующие классы:

- – **онтология** ПрО, которая описывает область, к которой относятся информационные потребности пользователя $O_{ПрО_i} = \left\langle T_{ПрО_i}, R_{ПрО_i}, F_{ПрО_i} \right\rangle, i = \overline{1,n}$;

- – **лексическая онтология** ПрО, которая содержит сведения о лексемах естественных языков, соответствующих терминам онтологии ПрО $O_{lex_i} = \left\langle T_{ПрО_{lexi}}, R_{lex_i}, f \right\rangle, i = \overline{1,n}$,

т.е. $\forall t \in T_{ПрО_i} \exists t_{lex} \in T_{ПрО_{lexi}}$, свойствами которого является множество лексем естественных языков, соответствующих данному термину онтологии, а отношения описывают связи между ними;

- **тезаурус** задачи– множество пар, первым элементом которых являются термины онтологии, совокупность которых характеризует ту конкретную задачу из ПрО, которую в данный момент решает пользователь, а вторым – вес (положительный или отрицательный) этого термина для данной задачи $Th_{i_j} = \left\{ \left\langle th_{k_{ij}} \in T_{ПрО_i}, v_{k_{ij}} \right\rangle, k = \overline{1, s_i} \right\}, j = \overline{1, m_i}$;

- **запрос** – множество ключевых слов, характеризующих одну их информационных потребностей пользователя, связанный с конкретной задачей при помощи тезауруса $z = \left\langle \{k_q\}, Th_{i_j} \right\rangle, q = \overline{1,u}$;

- **тема** – множество запросов, связанных с одной информационной потребностью $thema = \left\langle id_{thema}, \{z_q\} \right\rangle, q = \overline{1,u}$, которое может объединять запросы разных пользователей, базирующиеся на различных онтологиях и тезаурусах, и позволяющее объединять семантически связанные запросы;

- **результат запроса** – множество пар, первым элементом которых являются ссылки на ИР, а вторым – оценки этих ИР пользователем $rez = f(z,u) = \left\{ \left\langle id_{ir}, rating_{ir} \right\rangle \right\}$;

- **пользователь** u – класс, имеющий более сложную структуру и имеющий следующие атрибуты, которые можно разделить на несколько групп:

1. регистрационная информация:

- идентификатор пользователя;
- пароль для доступа к ИПС;

2. опыт взаимодействия ИПС с пользователем:

- список онтологий, которые пользователь применял для описания своих информационных интересов;
- список тезаурусов, которые пользователь применял в поисковых запросах;
- список ранее выполненных запросов;
- список результатов выполненных запросов с оценками пользователя для найденных результатов;

3. сведения, импортируемые из внешних

источников (необязательные сведения, могут отсутствовать):

- идентификаторы пользователя в социальных сетях, позволяющие динамически обновлять сведения о нем;
- рейтинги пользователя в социальных сетях;
- адрес пользователя в Википедии и других вики-ресурсах;
- адрес сайта пользователя;
- сфера компетенций пользователя (ключевые слова, импортируемые из социальных сетей);
- ссылки на публикации пользователя;

4. собственные характеристики пользователя:

- сфера компетенций пользователя (список ключевых слов, вводимых пользователем непосредственно);

5. формальные данные о пользователе (необязательные сведения, предоставляющие ИПС дополнительные сведения для формирования групп пользователей со схожими информационными потребностями):

- место жительства;
- возраст;
- профессия, образование и т.д.
- **группа пользователей** – класс, свойствами которого являются идентификатор группы и список пользователей, по тем или иным причинам объединенных в одну группу (группы могут формироваться явно путем выбора пользователя или автоматически на основе соответствия каким-либо условиям, например, группы пользователей со сходными формальными данными или выполняющих похожие запросы) $gr = \langle id_{gr_i}, \{u_i\} \rangle, i = \overline{1,n}$;
- **информационный ресурс** – сведения о найденных ранее ресурсах и их оценках $\left\langle U_{url}, \left\{ \langle z_i, m_i, q_i \rangle, i = \overline{1,n} \right\} \right\rangle$, включающие идентификатор ресурса, запросы, по которым он был обнаружен, оценку пользователя, которому он был предоставлен, и его уровень читабельности для этого пользователя.

Рекомендующие системы и семантический поиск

Рекомендующие системы (РС) [Ricci, 2011] отличаются от ИПС тем, что пользователю не надо явным образом формулировать поисковый запрос – система сама, на основании имеющихся сведений о пользователе, предлагает ему рекомендуемые элементы (РЭ). Чем выше и точнее информированность РС о потребностях пользователя, тем более эффективны результаты ее работы. Персонализированные рекомендации – это упорядоченные списки РЭ, т.е. работа РС сводится именно к ранжированию доступных РЭ.

При этом упорядочении РС, основываясь на предпочтениях и ограничениях пользователей, пытается прогнозировать, какие товары или сервисы наиболее подходят пользователю. Для этого РС накапливает информацию о предпочтениях пользователя и о его действиях. Эти подходы целесообразно использовать и при семантическом поиске, учитывая оценки, данные различным ИР различными сообществами пользователей, сгруппированными на основе подобных информационных потребностей [Рогушина, 2013]

Формально создание рекомендаций в РС может быть представлена следующим образом: пусть C – множество пользователей РС, S – множество предлагаемых РЭ (товаров, книг, фильмов, сервисов и т. д.). U – функция полезности, описывающая интерес пользователя $c \in C$ к РЭ $s \in S$, т. е. $U : C \times S \to R$, где R – количественная оценка. Цель РС – для каждого потребителя $c \in C$ выбрать такой РЭ $s` \in S$, что $U(c,s`) = \max_{s \in S} u(c,s)$. Каким именно образом определяется функция полезности, зависит от типа РС и от специфики РЭ.

При классификации РС обычно выделяют следующие подходы к отбору РЭ:

- *персональный* подход – анализ профиля конкретного пользователя, его ранее проявленных предпочтений и явным образом выраженных условий;
- *социальный* (коллаборативный) подход – анализ предпочтений других пользователей, которые по тем или иным причинам могут распространяться и на того пользователя, для которого делается выбор;
- *контент-ориентированный* подход, при котором анализируются сами РЭ, предлагаемые пользователю;
- *доверительный* подход – анализируется качество предлагаемых пользователю РЭ и анализируется степень доверия к ним.

Следует отметить, что в большинстве реальных РС все эти подходы реализуются интегрированно, но им придается различное внимание.

В большинстве РС подобие между двумя пользователями основывается на том, какие оценки они дали одним и тем же товарам. Наибольшее распространение получили корреляционный метод и метод линейного сходства.

Для эффективной работы РС надо предвидеть оценки, исходя из небольшого количества примеров.

Для преодоления проблему разреженности оценок следует при поиске похожих пользователей использовать также сведения из их профилей и обнаруживать пользователей со схожими профилями, например, относящихся к одному демографическому сегменту.

Анализ основных направлений развития современных РС [Middleton, 2009] связывает их с использованием онтологий для представления знаний как о пользователях, так и о РЭ. При персональном подходе РС необходимо накопить достаточно сведений о пользователе, чтобы в дальнейшем их обобщать и анализировать. Фоновый мониторинг работы пользователя обеспечивает положительные примеры того, что этот пользователь ищет, не мешая его нормальной работе. Для нахождения отрицательных примеров из наблюдаемого поведения также могут применяться эвристики, (хотя в целом с меньшей точностью). Эта идея лежит в основе тех РС, которые наблюдают за поведением пользователей и рекомендуют им те новые РЭ, которые коррелируют с профилями пользователей. Например, если пользователь регулярно просматривает сайты определенной тематики, то РС, проанализировав контент этих сайтов, может предложить ему другие сайты той же направленности.

Другой способ рекомендовать РЭ базируется на рейтингах, предоставляемых теми людьми, которые ранее оценили РЭ. Коллаборативные РС для этого запрашивают у пользователей явные оценки РЭ, а затем рекомендуют те РЭ, которые высоко оценили похожие пользователи.

Рекомендация относительно новых РЭ для пользователей может формироваться на основе его сравнения с подобными РЭ (фильтрация на основе контента), отзывов об РЭ в сообществе пользователей (коллаборативной фильтрации), семантических отношений между РЭ (эвристические рекомендации) или сочетания этих подходов. Во многих случаях выбор подхода зависит от того, насколько доступны метаданные об РЭ и есть ли обратная связь с пользователями (явно и неявно). Методы на основе контента хорошо работают, если есть достаточная обучающая выборка, а коллаборативные методы – когда система имеет большое сообщество пользователей. Однако на сегодня не выработаны общепринятые правила для выбора стратегии рекомендования.

Коллаборативная фильтрация использует рейтинги, предоставляемых сообществом пользователей, чтобы рекомендовать РЭ конкретному пользователю. Существуют два взаимодополняющих подхода к коллаборативной фильтрации: на основе пользователя или на основе РЭ. При коллаборативной фильтрации на основе пользователя находят группы подобных пользователей, а затем конкретному пользователю рекомендуют те РЭ, которые понравились другим пользователям из той же группы. При коллаборативной фильтрации на основе РЭ группируются те РЭ, которые одинаково оцениваются людьми. Для того, чтобы выполнить коллаборативную фильтрацию, должен быть создан профиль пользователя на основании имеющихся документов о том, какие РЭ были этим пользователем рассмотрены и оценены.

Так же, как и основанном на пользователе методе, сходство РЭ определяется при помощи того, сколько пользователей оценили эти РЭ как подобные. При этом определяются наборы похожих предметов. Этот метод хорошо масштабируется, поскольку новые РЭ добавляются к окрестностях на основе того, как пользователи оценивают их, без необходимости явного использования онтологии.

На основе вышеприведенного анализа можно предложить следующие подходы к работе РС с использованием онтологий: формирование модели пользователя; формирование модели РЭ; создание онтологии РЭ; накопление сведений об экземплярах РЭ и экземплярах пользователей; накопление оценок РЭ пользователями; анализ экземпляров РЭ; классификация (или кластеризация) пользователей на группы с подобными интересами; формирование набора стратегий, которые пользователь может явно выбирать для получения рекомендации; построение метода, позволяющего уточнить класс необходимого пользователю РЭ.

Широко распространенные классификации подходов к выработке рекомендаций, подразделяющие все существующие методы на базирующиеся на пользователе и базирующиеся на РЭ, а также на персональные и коллаборативные, являются слишком общими и, как правило, бинарными. Кроме того, в РС большое внимание уделяют алгоритмам вычисления подобия пользователей и РЭ, и значительно меньшее – методам классификации пользователей и РЭ.

На практике целесообразно не только использовать больше критериев, значимых для выработки рекомендаций, но и предоставить пользователю РС возможность самостоятельно формировать стратегию рекомендования, явным образом указывая значимость для данной задачи тех или иных критериев.

При выработке рекомендаций в РС многое зависит от специфики искомого РЭ.

Все РЭ можно подразделить на две категории с точки зрения возможности их *повторного использования*: используемые одноразово и многоразово. К первой категории относятся различные предметы материального мира и связанные с их использованием услуги. К ним относятся, например, технические устройства, продукты питания, авиабилеты, турпоездки. Если у пользователя уже есть такой РЭ, то ему может понадобиться такой же или похожий на него (после поломки первого, использования и т.д.). К второй категории относятся, как правило, информационные объекты, т.е. такие объекты, что наличие одного

экземпляра позволяет создавать произвольное количество его копий. К ним относятся, например, электронные книги и фильмы. Если у пользователя уже есть такой РЭ, то маловероятно, что ему понадобится еще один (хотя возможна утрата или поломка).

Кроме того, РЭ можно классифицировать (на два или более классов – в зависимости от требуемой точности рекомендаций) на *редко или часто* используемые. К примеру, прогноз погоды нужен пользователю почти ежедневно, а выбор модели холодильника актуален для большинства раз в 10 лет. Для наиболее часто используемых РЭ большинство пользователей склонны ориентироваться на собственный опыт, а для редко используемых – на совокупный опыт сообщества пользователей РС. При этом следует учитывать, что один и тот же РЭ может одним пользователям быть интересен редко, а другим – часто. Например, большинство людей редко интересуется особенностями газовых плит или мебельной фурнитурой, но специалист по комплексным ремонтам может выполнять такие запросы регулярно. Поэтому надо дать пользователю возможность самому явно определять, насколько часто он интересуется такими РЭ (т.е. насколько его собственное мнение о них компетентно).

Еще один важный параметр РЭ – *субъективность* оценивания. Если, к примеру, при оценке бытовой техники или автомобилей достаточно легко сформулировать те отличия, по которым пользователь более высоко оценивает один РЭ, чем другой (например, надежность работы, стоимость, простота обслуживания, функциональные возможности), и вследствие этого каждому пользователю может быть полезен опыт всего сообщества, то при оценке предметов искусства, музыки, фильмов такие отличия практически невозможно формализовать, и потому при выработке рекомендаций для конкретного пользователя важен только опыт некоторого подмножества сообщества с аналогичными вкусами, причем это подмножество может быть сформировано при помощи методов машинного обучения и на основе традуктивного вывода.

На способ выработки рекомендаций влияет и то, насколько оценивание РЭ требует *специальных знаний* в конкретной предметной области. Особенно важно это для тех РЭ, которыми большинство пользователей интересуются редко и потому сами, как правило, не имеют о них глубоких знаний. Вследствие этого они оценивают, как правило, лишь конкретный предмет (а не все предметы данного класса) и практически не имеют возможности сравнивать его с другими подобными РЭ. Например, оценивая телевизор, пользователь может оценить лишь ту модель телевизора, которую он купил, и сравнивать ее лишь с несколькими теми моделями, которыми он пользовался. Поэтому в некоторых областях важнее ориентироваться на мнение экспертов, а не на мнение большинства.

Открытым остается вопрос о том, каким образом формировать множество экспертов для той или иной ПрО. В частности, в большинстве социальных сетей существует как возможность зафиксировать как связь пользователя с набором тем, так и средства определения общего рейтинга пользователя (влиятельность его мнения для других пользователей, оценка его действий другими пользователями и т.д.), но, как правило, отсутствует возможность дифференцировать компетентность пользователя в той или иной тематике. Например, один и тот же пользователь, проявляющий интерес к компьютерам и кулинарии, может быть экспертом в компьютерной техники (и его оценки различных экземпляров компьютеров будут высоко точными), а в приготовлении пищи разбираться очень плохо (и именно потому и интересоваться этой областью) и потому оценивать различные РЭ крайне неправильно.

Следует отметить, что в процессе развития Web и систем электронной коммерции появилось много источников, обеспечивающих доступ к одним и тем же РЭ. При этом речь идет как о материальных, так и об информационных продуктах и услугах. Но при этом разные источники предлагают разные условия доступа и качества обслуживания. Поэтому значительного внимания заслуживает степень доверия к источникам, которую можно определить по совокупности оценок сообщества пользователей. Этот критерий рекомендования выделяется отдельно, так как речь идет не об оценивании РЭ, а об оценивании *источников РЭ*. Например, многие Web-сайты предлагают бесплатно скачивать статьи, книги и фильмы, но некоторые из них требуют регистрации, отправки SMS-сообщений (не бесплатных), оплаты пароля на распаковку архива и т.д. очевидно, что такие сайты оцениваются значительно ниже, чем те, которые предлагают свободный доступ к электронной библиотеке. Электронные магазины, предлагающие покупку материальных предметов, также работают не всегда честно – они могут задерживать доставку, предоставлять бракованную продукцию, требовать дополнительной оплаты доставки, значительно завышать цену по сравнению с обозначенной (например, на сайте цена представлена в долларах, а оплату следует осуществлять в гривнах по крайне невыгодному курсу, отличающемуся от официального). Поэтому следует предоставить пользователям оценить такие магазины ниже, чем те, которые работают корректно.

В большинстве используемых на практике РС вырабатываются рекомендации относительно какого-то довольно узкого класса РЭ, и нет возможности связывать профили одного и того же пользователя, сформированные различными РС (например, нельзя связать предпочтения пользователя относительно выбора художественной литературы, покупок в электронном магазине и при просмотре новостей). Использование онтологических моделей пользователей позволяет в какой-то мере решить эту задачу и интегрировать

различные РС. При этом возникает дополнительная задача – классификация РЭ, относительно которого пользователь нуждается в рекомендации (и, соответственно, переадресация запроса к соответствующей специализированной РС). Следует учитывать, что нередко пользователь нуждается в рекомендациях по набору взаимосвязанных вопросов (например, выбор места для отдыха связан и с выбором турпутевки, и с прогнозом погоды, и с рекомендациями относительно транспорта). Для этого необходимо разработать общую онтологию РЭ, которая должна быть достаточно компактной и несложной, но при этом охватывать основные классы РЭ, относительно которых пользователи часто нуждаются в рекомендациях. Специфичные знания предметных областей такая онтология не должна включать, т. к. они должны содержаться в онтологиях специализированных РС.

Рассмотрим также, оценки каких именно групп пользователей целесообразно применять для коллаборативной фильтрации. Самый простой случай группы – это группа, состоящая всего из *одного пользователя*, для которого и осуществляется поиск рекомендаций.

Можно сказать, что при этом коллаборативная фильтрация сводится к персональной. Противоположный случай – когда для пользователя значимы оценки РЭ *всем сообществом* в целом. Это может иметь место для тех ПрО, к которым пользователь обращается впервые и еще не имеет собственного мнения не только о самой области, но и о критериях нахождения в ней экспертов. К промежуточным случаям относятся анализ оценок экспертов в ПрО (целесообразно предоставить пользователю возможность явно задавать приемлемый уровень их квалификации).

В целом следует оценивать выбранную стратегию рекомендования по трем направлениям – учет мнения самого пользователя, учет мнения сообщества, анализ самого РЭ. В таком трехмерном пространстве можно разместить большинство типичных объектов рекомендования.

Выбор пользователем значения по каждому из трех параметров для стратегии рекомендования для интересующих его РЭ и наличие онтологии (или хотя бы таксономии РЭ) позволяет достаточно точно профилировать интересы самого пользователя, оценить его собственную компетентность для оценивания РЭ (и, соответственно, значимость его мнения для других пользователей) и выявлять группы пользователей со сходными интересами. Следует отметить, что более объективным является признание своей некомпетентности в оценивании РЭ, чем декларирование своей высокой квалификации.

Тезаурусы как средство представления знаний

Следует отметить, что важным требованием пользователя к работе любой информационной системы является понятность и предсказуемость ее действий. В общем случае онтология ПрО – достаточно сложная структура, а конкретная задача, для решения которой пользователь ищет информацию, использует только часть содержащихся в такой онтологии знаний. Поэтому представляется целесообразным использовать для моделирования знаний пользователя об интересующей его ПрО частного случая онтологии – тезауруса, который можно рассматривать как проекцию онтологии на задачу.

Тезаурус можно представить как комплекс лингвистических знаний [Браславский, 1997]. Сегодня именно с тезаурусами связаны многие направления усовершенствования семантического поиска [Лукашевич, 2011].

Тезаурус в ИТ – это полный систематизированный набор данных о какой-либо области знаний, позволяющий человеку или вычислительной машине в ней ориентироваться. Тезаурус – это $Ts = \langle T, R \rangle$, где T – множество терминов, а R – множество отношений между этими терминами. Множества T и R конечны. Множество терминов тезауруса T соответствует множеству концептов X онтологии O. Тезаурусы позволяют моделировать знания как о пользователях, так и о тех ресурсах, которые они ищут [Гладун, 2006].

Чтобы формализовать область своих интересов – ПрО поиска – пользователь должен создать тезаурус, моделирующий интересующую его ПрО, в котором содержатся основные термины ПрО и связи между ними. Тезаурус можно создать вручную или автоматизированно. Основой для автоматического создания тезауруса может послужить обработка набора релевантными этой ПрО или ранее созданная онтология ПрО, из которой пользователь отбирает только необходимые ему термины. Все эти подходы могут комбинироваться друг с другом.

Для создания тезаурусов ИР и РЭ предлагается использовать упрощенный алгоритм построения тезауруса: по полному перечню слов, используемых в ИР, строится словарь терминов, из которого отбрасываются стоп-слова, содержащиеся в специально разработанном пользователем списке. Этот алгоритм применяется только для тех ИР, которые не сопровождаются метаописаниями. В противном случае из метаописаний (в формате RDF или OWL) извлекаются термины тезауруса и связи между ними, которые дополняют построенный по контенту ИР словарь. Аналогично строятся тезаурусы РЭ – обрабатываются их метаописание, контент, отзывы о них других пользователей.

Пользователь вводит запрос, приблизительно идентифицируя свою информационную потребность с помощью ключевых слов или выбирая класс интересующего его РЭ (возможно, с набором условий и ограничений), например, РЭ

класса «художественная литература/фантастика/фэнтези», изданная после 2005 года. В ответ РС формирует набор РЭ, доступных системе и соответствующих этому приблизительному запросу – n ссылок на РЭ и их кратких описаний $I=\{Ref_j,D_j\}$, $j=\overline{1,n}$. Здесь Ref_j – ссылка на соответствующий РЭ (или его описание), а d_j – информация об этом РЭ, доступная РС.

Если множество I не пусто, причем РС найден в ответ на запрос более чем один РЭ ($n\ge 1$), то нужно установить порядок, в каком предлагать пользователю сведения о найденных РЭ. Тогда для всех РЭ из этого множества $I=\{Ref_j,D_j\}$, $j=\overline{1,n}$ формируются их упрощенны тезаурусы $Ts(ИР_j)=\langle T_{j_j},\varnothing\rangle$, $j=\overline{1,n}$ и соответствующие им словари терминов $T_j=\{t_{j_{1w}}\}$, $j=\overline{1,n}$, $w=\overline{1,q_j}$. $t_{j_{lw}}$ – это слова, которые используются в информации о j-м РЭ, найденном РС, т. е. в $D_j, j=\overline{1,n}$. $q_j, j=\overline{1,n}$ – это количество различных слов, используемых в описании $D_j, j=\overline{1,n}$. Если слова в описании повторяются, то в словаре терминов они фиксируются только один раз.

Затем пользователь формирует тезаурус интересующей его ПрО (или указывает на ранее сформированный тезаурус) $Ts_{ПрО}$ и соответствующий ему словарь терминов этой ПрО $T_{ПрО}=\{t_m\}$, $m=\overline{1,q}$. $T_{ПрО}$ – это множество, состоящее из m терминов, относящихся к интересующей пользователя ПрО. Это множество строится аналогично словарю терминов РЭ и обычно формируется как объединение словарей терминов, содержащихся в документах, которые пользователь нашел ранее и посчитал релевантными интересующей его ПрО (как в их контенте, так и в метаописаниях).

Производится сравнение $T_{ПрО}$ и T_j, $j=\overline{1,n}$, высчитывается коэффициент их близости

$$K_j=\sum_{m=1}^{q}\sum_{w=1}^{w_j} f(t_{j_w},t_m),\ m=\overline{1,q},\ w=\overline{1,w_j},$$

где

$$f(t_1,t_2)=\begin{cases}0, & \text{если } t_1\ne t_2,\\ 1, & \text{если } t_1=t_2.\end{cases} \qquad (1)$$

Коэффициент (1) представляет собой количество терминов, которые встретились как в тезаурусе РЭ, так и в тезаурусе ПрО. Найденные ИР упорядочиваются в зависимости от значений K_j, пользователю предъявляются в первую очередь те ИР, которые имеют наиболее высокий коэффициент близости к ПрО.

При использовании коэффициента (1) возникает следующая проблема: слова, соответствующие одному термину, но являющиеся, например, различными словоформами, синонимами или переводами на различные языки, обрабатываются как разные термины. Поэтому представляется целесообразным использовать онтологию ПрО и выделять группы слов, соответствующих одному термину. Для этого пользователь должен связать РЭ словаря терминов тезауруса ПрО с одним из терминов онтологии ПрО, т. е. $\forall t_m\in T_{ПрО}$, $m=\overline{1,q}$ задать функцию $g(t_m)\in X$. Затем для вычисления коэффициента близости K^O эта функция используется следующим образом:

$$K^O{}_j=\sum_{m,l} f(t_{j_w},t_m),\ m=\overline{1,q},\ w=\overline{1,w_j},$$

где $f(t_1,t_2)=\begin{cases}0, & \text{если } g(t_1)\ne g(t_2),\\ 1, & \text{если } g(t_1)=g(t_2).\end{cases}$ (2)

Коэффициент (2) представляет собой количество терминов, которые встретились как в тезаурусе РЭ, так и в тезаурусе ПрО и при этом ссылаются на один и тот же термин онтологии ПрО. По сравнению с коэффициентом (1) коэффициент (2) позволяет использовать меньший объем документов для построения тезауруса ПрО, но требует большее время для вычислений.

При создании тезауруса ПрО, которая интересует пользователя РС, необходимо явно указать основные понятия ПрО и связи между ними. К сожалению, большинству пользователей достаточно сложно это сделать (даже имея соответствующие знания и применяя их в своей деятельности). На первом этапе формирования тезауруса пользователь может выбрать одно из следующих решений: 1) самостоятельно построить с помощью одного из редакторов онтологий онтологическое описание интересующей его ПрО; 2) найти (например, в Web) онтологию на языке OWL, которую описывает ПрО, близкую к области его информационных интересов; 3) сформировать множество понятий ПрО, содержащее наиболее характерные слова и словосочетания, встречающиеся в интересующих его ИР.

Использование индуктивного вывода для пополнения онтологий

При построении онтологий и тезаурусов ПрО, а также при коллаборативном подходе к поиску важно определить, какие связи между терминами ПрО являются существенными для описания информационной потребности пользователя. Пользователю достаточно сложно самостоятельно обнаружить все важные закономерности и отбросить несущественные.

Для их выявления можно воспользоваться методами индуктивного и традуктивного извлечения знаний из данных. Существуют независимые подходы к реализации подобных методов: ID3, ACLS, CART и т. д. Наиболее интересным, в связи со спецификой проводимой работы, оказался алгоритм ID3 [Quinlan, 1979], который специально разработан для извлечения ценной информации из больших объемов слабо структурированных данных. При работе этого алгоритма время вычислений зависит линейно от числа введенных примеров, числа атрибутов, используемых для описания примеров, и числа узлов в строящемся дереве решений, что отличает его от таких известных алгоритмов построения деревьев решений, как INDUCE, SPROUTER, ROTH-P, в которых усилия, требующиеся для решения задачи, резко возрастают вместе со сложностью задачи.

Если методы, подобные МГУА (метод группового учета РЭ), предназначены для нахождения закономерностей по набору количественных измерений параметров и полученному по ним результату, то методы, подобные ID3 и его вариациям (C4.5, ID4 и т.д.), предназначены для обобщения опыта экспериментов, параметры и результаты которых описаны через качественные оценки (лингвистические переменные). В большинстве случаев между их значениями невозможно установить даже относительное упорядочение (например, различные симптомы и диагнозы пациентов). К таким задачам относится и проблема, которую решают рекомендующие системы. ID3 принадлежит к невозрастающим алгоритмам, то есть при добавлении к набору классифицированных примеров определенного количества новых нужно обрабатывать снова как старые, так и новые примеры. Но ID3 предназначен для построения только бинарного дерева решений, а этого недостаточно удобно для представления закономерностей многих ПрО.

Поэтому для пополнения онтологий предлагается использовать *ID3m* [Rogushina, 2012] – модификацию ID3 для произвольного (конечного) количества решений. Он также принадлежит к невозрастающим алгоритмам. В данном случае, примерами обучающей выборки являются РЭ, доступные РС, а параметрами, по которым они описываются, являются их свойства, описанные в метаданных и в онтологии РЭ, а также термины тезауруса пользователя. Значения, соответствующие терминам тезауруса, – "Термин отсутствует в описании РЭ", "Термин встречается в описании РЭ редко", "Термин встречается в описании РЭ часто". В качестве результата используется оценка, данная пользователем найденному РЭ (качественная оценка, имеющая два и более значений).

На вход алгоритма поступает обучающая выборка H – набор из n классифицированных (получивших одну из возможных оценок) примеров одинаковой размерности $H = \{h_i\},\ i = \overline{1,n}$.

Каждый пример из выборки представляет собой упорядоченную последовательность значений s атрибутов и результирующего атрибута $h_i = \langle a_1, \ldots a_s, r \rangle$, $i = \overline{1,n}$. Значения атрибутов принадлежат конечным множествам: $a_{j_u} \in A_j$, $j = \overline{1,n}$, $u = \overline{1,n_j}$, $r_y \in R$, $y = \overline{1,n_r}$.

Если обучающая выборка содержит примеры, в которых все значения атрибутов одинаковы, а решения различны, то введенная информация недостаточна для построения классификационного правила. Если множество примеров пустое, то можно произвольно связать его с любым решением. Если все примеры относятся к одному классу, строится один лист дерева решений, связанный с этим классом. В противном случае необходимо выбрать один из атрибутов и разделить множество атрибутов на подмножества в зависимости от значения этого атрибута и применить алгоритм к каждому из полученных подмножеств.

На каждом шаге работы алгоритма вычисляется, какой атрибут m несет наибольшее количество информации о результате.

$$C_{max} = \max\left\{C_z,\ z = \overline{1,s}\right\} =$$

$$= \max_z \left\{ \sum_i \sum_j \frac{C\left(a_{z_i} \in A_z, r_j \in R_j\right)}{d_z} \right\}, \quad (3)$$

где C(x,y) – количество информации, d_m – стоимость получения значения m-го атрибута. В результате работы алгоритма ID3m формируется дерево решений, в котором каждый лист связан с одним из решений, каждый узел – с именем одного из атрибутов, а выходящие из такого узла ветви – со значениями этого атрибута.

Такое дерево решений позволяет РС по параметрам вновь найденного РЭ прогнозировать, как именно оценит его пользователь, и предлагать пользователю в первую очередь те РЭ, которые соответствуют его индивидуальным предпочтениям. Так как точные значения вероятностей событий из обучающей выборки неизвестны, то они аппроксимируются на основе рассматриваемого множества примеров.

Предложенный выше подход к формированию рекомендаций основывается на использовании знаний пользователя о ПрО, характеризующей его информационные потребности. Пользователь может явно указывать интересующие его термины и получать те РЭ, которые соответствуют его потребности. Такой подход ориентирован на пользователя с относительно стабильными информационными потребностями, не являющегося специалистом в области информационных

технологий, и позволяет пользователю избежать рутинной работы по фильтрации результатов поиска в Web.

МАИПС и выработка рекомендаций

МАИПС – мультиагентная информационно-поисковая система с развитыми средствами интеллектуализации ее поведения, предназначенная для поиска информации в описанных пользователем относительно узких предметных областях, связанных с профессиональными или научными интересами пользователей, и рекомендует пользователю те результаты поиска, которые относятся к интересующей его ПрО и соответствуют его информационным потребностям. Ее можно рассматривать как рекомендующую систему, ориентированную на формирование рекомендаций относительно естественно-языковых и мультимедийных информационных ресурсов, доступных через Web.

Система МАИПС ориентирована на пользователей, имеющих в сети постоянные информационные интересы и требующих постоянного поступления соответствующей информации. Для этого МАИПС позволяет сохранять и повторно выполнять запросы, учитывая реакцию пользователя на ранее предложенные ему ИР (персональная фильтрация), отслеживать появление аналогичных запросов у других пользователей (коллаборативная фильтрация), сохранять формальное описание области интересов пользователя в виде онтологии (семантическая фильтрация) и т.д. Кроме того, в МАИПС при профилировании пользователей используется специфичный для естественно-языковых ИР критерий оценивания – сложность текста для понимания. Особенностью системы является использование оригинального знание-ориентированного алгоритма, позволяющего определить сложность понимания текста для конкретного пользователя (для этого используются тезаурусы предметных областей, интересующих пользователей) [Рогушина, 2007].

Важной особенностью МАИПС является то, что в этой системе интегрированы различные средства и методы создания, пополнения, обработки и использования онтологий, которыми пользователь может управлять явным образом. Для представления знаний об интересующей пользователя ПрО используются онтологии и тезаурусы ПрО: тезаурус строится пользователем по соответствующей онтологии самостоятельно, а онтология выбирается из набора предложенных на сайте либо импортируются из Web.

Пользователь МАИПС может обращаться к онтологиям, созданных другими пользователями – пересматривать их, задавать по ним контекст поиска, копировать из них нужны фрагменты, но не имеет права изменять их. ИПС может обеспечить поиск онтологий, которые содержат введенные пользователем термины, а также поиск онтологий, похожих на выбранную пользователем онтологию. Это позволяет создавать группы пользователей с общими информационными интересами и предотвратить дублированию в выполнении одинаковых многоразовых запросов разных пользователей.

Основой МАИПС являются технологии Semantic Web, в частности, язык представления онтологий OWL DL и средства его обработки (рис.2).

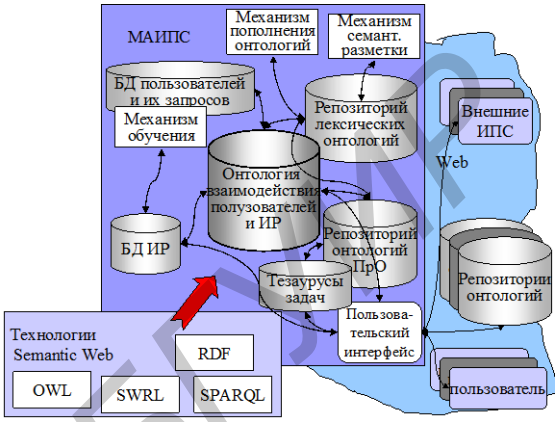

Рисунок 2 - Использование онтологической модели взаимодействия пользователей и ИР для семантического поиска в МАИПС

По мере развития МАИПС возникла потребность в подключении репозитория онтологий, чтобы пользователи могли повторно использовать знания о ПрО, доступные в Web. При этом поиск может осуществляться не только по ключевым словам, а и по другим важным свойствам онтологий. Поэтому в дальнейшем представляется целесообразным реализовать в МАИПС средства взаимодействия с репозиториями онтологий, поддерживающие поиск нужной пользователю онтологии, обнаружение похожих на выбранную пользователем онтологий, а также сопоставление построенного пользователем тезауруса с другими онтологиями и тезаурусами.

Онтологическая модель, описывающая семантику взаимодействия пользователей и ресурсов МАИПС в информационном пространстве Web, обеспечивает знания для выполнения следующих действий, связанных с поиском информации и основанные на рассмотренных выше методах:

1. *предварительный этап,* когда в систему вводятся сведения о окружающем мире, то есть

- создается онтологическая модель, описывающая структуру информации относительно основных элементов, с которыми работает система (пользователей, ресурсов, результатов поиска и т.д.) (рис.1);

- вводятся онтологии ПрО, которые могут быть полезны для поиска, и ссылки на внешние репозитории и средства поиска онтологий, которые

пользователь может применить для работы в более специфических ПрО;

- по имеющимся онтологиями создается несколько примеров тезаурусов, которые могут быть использованы при поиске;

2. этап *регистрации пользователя*, на котором пользователь вводит сведения, необходимые для создания нового экземпляра класса ~~п~~ользователь";

3. этап *создания нового запроса* пользователя, на котором пользователю последовательно нужно выполнить следующие действия :

- Выбрать базовую онтологию, знания которой обеспечат семантическую обработку запроса;

- По выбранной онтологией создать тезаурус запроса одним из следующих способов:

* выбрать несколько классов или экземпляров классов из базовой онтологии;

* выбрать несколько классов из базовой онтологии и классы, находящиеся от них на семантической расстоянии не более указанной пользователем величину;

* выбрать несколько классов из базовой онтологии и их надклассы и подклассы указанной пользователем глубины;

* выбрать несколько классов из базовой онтологии и классы, связанные с выбранными классами выбранным пользователем отношением, специфичным для ПрО;

* вручную ввести термины тезауруса, характеризующие интересующую пользователя задачу;

* над построенными ранее тезаурусами применить теоретико-множественные операции объединения, пересечения и дополнения;

* указать вес каждого из терминов тезауруса, который отражает его важность для конкретной задачи.

- создать список ключевых слов, характеризующих конкретный информационный запрос, и соединить его с одним из ранее созданных тезаурусов;

- если нужно, присоединить запрос к одной из ранее построенных групп запросов или создать для него новую группу.

4. этап *выполнения запроса*, в процессе которого поисковый запрос по ключевым словам перенаправляется внешней ИПС (Google), затем МАИПС получает найдены результаты и переупорядочивает их в соответствии с количеством найденных в них терминов тезауруса и их весом.

Кроме того, для упорядочения могут учитываться другие свойства ИР, например, если пользователь указывает желаемый уровень читабельности, то этот параметр тоже влияет на рейтингование ИР.

Если некоторые из найденных ИР ранее были предложены другим пользователям МАИПС, то по желанию того пользователя, который предоставляет запрос, их оценки могут учитываться либо непосредственно, либо с учетом таких факторов, как степень подобия между этими пользователями и их рейтинг в данной ПрО, который вычисляется как по их собственным оценкам и сведениям о них из социальных сетей, так и по статистике, накопленной МАИПС.

5. этап *обработки запросов* для повышения эффективности поиска. Кроме непосредственного выполнения запросов МАИПС выполняет следующие функции:

- определения уровня компетентности пользователей для различных ПрО, соответствующие имеющимся онтологии ПрО, которое базируется на таких параметрах, как :

* Количество запросов пользователя, основанных на данной онтологии;

* Собственная оценка пользователем своей осведомленности в интересующих его ПрО;

* Выбранный пользователем уровень читабельности текстовых IP для данной ПрО - как заданный явно, так и средний для избранных им ИР по запросам, базирующихся на данной онтологии (этот параметр является наиболее объективным)

* Релевантность собственных публикаций пользователя этой ПрО, которая определяется путем сопоставления тезауруса, построенного по онтологией ПрО, с тезаурусами публикаций;

* Рейтинг пользователя в социальных сетях;

* Количество других пользователей МАИПС, выбирающих пользователя за эксперта для поиска рекомендаций в данной ПрО.

6. *создание рекомендаций* по результатам обработки. В отличие от большинства существующих рекомендующих систем, МАИПС позволяет пользователю явно, непосредственно и динамично управлять средствами создания рекомендаций.

Пользователь может учитывать оценки:

* Всего сообщества пользователей МАИПС;

* Подмножеств пользователей, запросы которых базируются на тех же онтологиях;

* Тех пользователей, которые используют наиболее подобные тезаурусы и ключевые слова для запросов;

* Запросов с выбранной пользователем темы, в которую могут входить как только собственные запросы пользователя, так и запросы различных пользователей с различными онтологиями и тезаурусами;

* Явно указанной подмножества пользователей МАИПС;

* Подмножества пользователей МАИПС, построенной по введенным пользователем формальными условиями (например, по месту жительства или по возрасту);

* Подмножества запросов самого пользователя, отвечающих определенным условиям (например, построенные в указанный интервал времени и по использованию ключевого слова);

Такие условия пользователь может предоставлять как для каждого запроса по отдельности, так и для определенной группы своих запросов. Результаты построения рекомендаций 1) влияют на упорядочение результатов запроса 2) позволяют рекомендовать пользователю те ИР, которые не встречаются в результатах его собственных запросов, но считаются МАИПС интересными для него

7. автоматизированное *пополнение профиля*. МАИПС функционирует в открытом информационном среде Web и поэтому нуждается в динамическом обновлении своих знаний об этом среду. Для этого используется проактивный поиск новых сведений о своих пользователях в Web (например, поиск их новых публикаций или тех публикаций, содержащих ссылки на них, экспорт новых сведений из социальных сетей) и автоматизированное обновление онтологий ПрО (например, путем обработки семантических вики-ресурсов, экспорта внешних онтологий).

8. *семантическая разметка* предоставляемых пользователю ИР, которая выполняется автоматически на основе лексической онтологии, соответствующей выбранной пользователем онтологии ПрО. В разметке, в зависимости от желания пользователя, могут применяться как все термины, присутствующие в этой онтологии, так и некоторое их подмножество – например, только те, которые присутствуют и в онтологии, и в построенном на ее основе тезаурусе задачи, или же еще и те, которые связаны с терминами из тезауруса выбранными пользователем отношениями.

Заключение

Совместное использование формальной онтологической модели взаимодействия пользователей с ресурсами и методов выработки рекомендаций, доступа к внешним источникам информации, индуктивного извлечения знаний и технологий Semantic Web при семантическом поиске в Web позволяет более эффективно обеспечить пользователя необходимыми сведениями, а явным образом выбранные методы рекомендования и онтологические описания ПрО обеспечивают пользователю понимание поведения такой системы.

Библиография

[Amerland, 2013] Amerland D. Google Semantic Search: Search Engine Optimization (SEO) Techniques That Gets Your Company More Traffic, Increases Brand Impact and Amplifies Your Online Presence. – Que Publishing, 2013. – 230 p.

[Middleton, 2009] Middleton S., De Roure D., Shadbolt N. Ontology-Based Recommender Systems // in Handbook on Ontologies, Edt. by S.Staab, R.Studer, Springer, 2009. – P. 779-796.

[Quinlan, 1979] Quinlan J.R. Discovery rules from large collections of examples: a case study // Expert Systems in the Microelectronic Age. – Edinburg, 1979. – P. 87-102.

[Ricci, 2011] Ricci F., Rokach L., Shapira B., Kantor P. Recommender Systems Handbook. – Springer, 2011. – 842 p.

[Rogushina, 2012] Rogushina J., Gladun A. Ontology-based competency analyses in new research domains // Journal of Computing and Information Technology. V.20, N. 4, 2012. – P.277-293.

[Бездушный, 2003] Бездушный А.Н., Гаврилова Э.А., Серебряков В.А., Шкотин А.В. Место онтологий в единой интегрированной системе РАН // Современные технологии в информационном обеспечении науки, М.: Научный Мир, 2003. С.97-115.

[Браславский, 1997] Браславский П.И., Гольдштейн С.Л., Ткаченко Т.Я. Тезаурус как средство описания систем знаний // Информационные процессы и системы. – 1997. – № 11, Серия 2. – С. 16-22.

[Гладун, 2006] Гладун А.Я., Рогушина Ю.В. Онтологии и мультилингвистические тезаурусы как основа семантического поиска информационных ресурсов Интернет // The Proc. of XII-th Intern. Conf. KDS'2006, Varna, Bulgaria. – P. 115-121.

[Лукашевич, 2011] Лукашевич Н.В. Тезаурусы в задачах информационного поиска. - М.: Издательство Московского университета, 2011. - 512 с.

[Рогушина, 2007] Рогушина Ю.В. Использование критериев оценки удобочитаемости текста для поиска информации, соответствующей реальным потребностям пользователя // Проблемы программирования. – 2007. –№ 3. – С. 76-87.

[Рогушина, 2013] Рогушина Ю.В. Использование онтологических знаний в рекомендующих системах // Проблеми програмування, 2013, №2. – С.71-86.

KNOWLEDGE-ORIENTED MEANS OF SEMANTIC SEARCH INTO THE WEB

Rogushina J.

Institute of software systems of National Academy of Sciences Ukraine, Kiev, Ukraine

ladamandraka2010@gmail.com

Semantic-based approach for Web search is proposed. Means of knowledge representation (ontologies and thesauri), methods of their refinement and matching are analysed. Ways of personal and collaborative information use are proposed.

Keywords: semantic search, ontological model, thesaurus, inductive inference, recommending systems, collaborative search.