РЕШЕНИЕ ЗАДАЧИ ПОИСКА ИНФОРМАЦИИ НА ОСНОВЕ ОНТОЛОГИЙ

Д.Е. Пальчунов,
д.ф.–м.н., профессор, заведующий кафедрой Общей информатики Новосибирского государственного университета, ведущий научный сотрудник Института математики СО РАН.

Рассмотрены методы информационного поиска и проблема поиска информации в Интернете. Проанализированы преимущества и недостатки известных поисковых систем; разработана метапоисковая система с интерфейсом в виде виртуального каталога. Предлагаемый подход основан на применении онтологий предметных областей. В работе использована теоретико-модельная формализация онтологий. Для поиска информации в Интернете применены иерархия онтологий предметных областей, онтология Интернет-ресурсов и онтология пользователя.

1. ВВЕДЕНИЕ

Работа посвящена методам поиска информации в корпоративных информационных системах и в Интернете[1]. Наибольший интерес представляет поиск информации, представленной различными Интернет-ресурсами. Это огромный объём информации, во многих случаях являющейся исчерпывающей. Организация поиска в корпоративных информационных системах имеет много общего с организацией поиска в Интернете. Принципиальная разница только в объёме информации и в наличии большого количества уже существующих поисковых систем для Интернета (Гугл, Яндекс и др.).

Наш подход основан на применении онтологий [8, 9, 28]. Для решения проблемы информационного поиска в Интернете использованы три вида онтологий: иерархия онтологий предметных областей; онтология Интернет-ресурсов и онтология пользователя.

Иерархия онтологий предметных областей содержит онтологии разделов и подразделов некоторой предметной области. Мы рассматриваем три таких предметных области – математику, катализ и патентоведение, представленными иерархиями подобластей. Математика состоит из алгебры, логики, анализа, геометрии, топологии, теории вероятностей и т.д.; алгебра – из теории групп, теории колец, теории чисел и т.д. Каждой подобласти (на каждом уровне) соответствует её онтология. В результате получаются иерархии онтологий для предметных областей – для математики, катализа и патентоведения.

Использование только иерархии онтологий предметных областей считаем недостаточным, поскольку даже по узкой области знаний в Интернете десятки тысяч ресурсов. С другой стороны, пользователю, как правило, нужны не все ресурсы, а только какой-то их вид: текст статьи, страница конференции, форум, персональная страница и т.п. Для более точной спецификации запроса пользователя необходимо использовать онтологию Интернета, описывающую виды и подвиды различных Интернет-ресурсов. И, наконец, для наиболее точного формулирования поискового запроса полезно определить вид задачи, решаемой пользователем, – получить ответ на некоторый вопрос, скачать статью, фильм или фотографию, купить книгу и т.д. Поэтому необходима онтология пользователя, полезная также для кастомизации – подстройки поисковой системы под конкретного пользователя.

Для реализации предлагаемого подхода нами использована теоретико-модельная формализация онтологий [8, 28]. Онтология предметной области рассматривается как пара – сигнатура из множества ключевых понятий предметной области и множество аналитических предложений, истинных в данной

[1]*«Работа выполнена при поддержке гранта РФФИ 05-01-04003-НИИО-а (DFG project COMO, GZ: 436 RUS 113/829/0-1), а также гранта Междисциплинарного интеграционного проекта СО РАН № 115 "Разработка интеллектуальных информационных технологий генерации и анализа знаний для поддержки фундаментальных научных исследований в области естественных наук"»*

предметной области. Это множество аналитических предложений определяет смысл (значение) ключевых понятий предметной области.

Для организации поиска научно-технической информации в Интернете разработана метапоисковая система, интерфейс которой реализован в виде виртуального каталога. Достижение точности и полноты информационного поиска в этой метапоисковой системе обеспечено методами формулировки поискового запроса, основанными на использовании онтологии предметной области, онтологии сети Интернет и онтологии пользователя.

2. ПОИСК ИНФОРМАЦИИ В СЕТИ ИНТЕРНЕТ

Организация точного поиска информации в сети Интернет – одна из наиболее бурно развивающихся областей в инженерии знаний [1, 2, 6, 9, 10, 13–15, 29]. Наиболее массовые инструменты информационного поиска – поисковые системы и Интернет-каталоги.

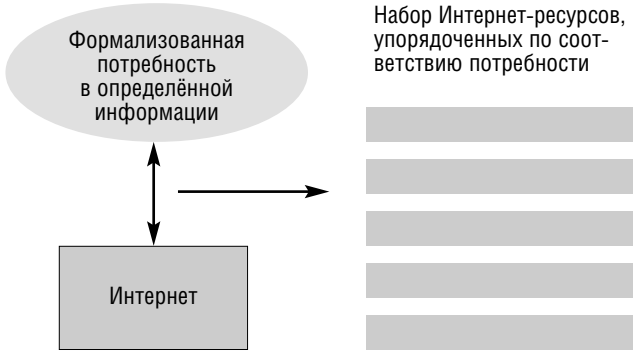
Информационно-поисковые системы позволяют вести поиск ресурсов по запросу пользователя, сформулированному в виде последовательности слов, а также предоставляют возможность расширенного поиска: требовать обязательного присутствия слов, отсутствия слов, использовать логические связки «и» и «или». Расширенный запрос – достаточно простая булевая комбинация утверждений вида «данное слово встречается в тексте».

Интернет-каталог состоит из структурированного набора Интернет-ресурсов, разбитых на рубрики и подрубрики – обычно несколько уровней вложенности. Все ссылки, в каталоге привязаны к этим рубрикам. Для поиска требуемого Интернет-ресурса нужно выбрать подходящую рубрику и просмотреть список относящихся к ней ссылок.

Информационно-поисковые системы и Интернет-каталоги имеют свои преимущества и недостатки. Чтобы оценивать качество поиска в различных информационно-поисковых системах и каталогах необходимо ввести характеристики информационного поиска.

Задачу информационного поиска в сети Интернет можно представить в виде приведённого ниже рисунка.

Пользователь для удовлетворения определённой информационной потребности формализует её средствами, которые ему даёт тот или иной поисковый инструмент. Если пользователь имеет дело с информационно-поисковой системой Гугл или Яндекс, информационная потребность будет формализована в виде последовательности слов или в виде расширенного запроса. Если поиск происходит по Интернет-каталогу, формализацией информационной потребности будет выбранная рубрика каталога.



После того, как информационная потребность пользователя формализована, происходит поиск. Пользователь получает список Интернет-ресурсов, соответствующих его запросу. Информационно-поисковые системы используют для этого свой индекс – набор проиндексированных ранее Интернет-ресурсов. Интернет-каталог выдаёт все хранящиеся в нём ссылки, относящиеся к выбранной рубрике. Выдаваемый список Интернет-ресурсов упорядочен по степени соответствия поисковому запросу. Таким образом, в нашей модели информационного поиска имеем: пользователя, желающего удовлетворить определённую информационную потребность, формализованный запрос к поисковой системе и выдачу поисковой системы – упорядоченный набор Интернет-ресурсов. Исходя из этой модели, сформулируем параметры качества информационного поиска. Наиболее распространённые параметры качества: релевантность, пертинентность, точность, полнота.

Пертинентность – мера качества поиска; насколько хорошо результат поиска удовлетворяет информационную потребность пользователя, т.е. это соответствие информации, полученной в результате поиска, потребности пользователя. Пертинентность определяется субъективным восприятием пользователя: в какой степени документ удовлетворяет его информационную потребность. Информационная потребность пользователя может быть выражена в формализованном запросе с той или иной степенью полноты и точности.

Релевантность – мера того, как хорошо список результатов отвечает на запрос; определяет порядок, в котором результаты поиска представлены пользователю. Когда есть большое количество найденных ресурсов, поисковая машина сортирует список результатов так, чтобы более релевантные страницы оказались в списке раньше, чем менее релевантные.

Релевантность – более узкое понятие, чем пертинентность. Документ может быть релевантен формализованному запросу, но при этом может

не удовлетворять информационную потребность пользователя. Различают содержательную и формальную релевантность – по методу её определения. Формальная релевантность – соответствие, определяемое алгоритмически сравнением поискового предписания и поискового образа документа, на основании применяемого в информационно-поисковой системе критерия выдачи. Содержательная релевантность – соответствие документа информационному запросу, определяемое неформальным путем.

Пертинентность – степень соответствия между ожиданиями пользователя и результатами поиска, отношение объёма полезной для пользователя информации к общему объёму полученной информации, найденной поисковой системой. Достижение высокой пертинентности – основная задача современных поисковых систем.

Существуют еще две характеристики информационного поиска, более глубоко раскрывающие релевантность и пертинентность.

Точность – мера эффективности поиска, выраженная в виде отношения числа найденных релевантных ресурсов к общему количеству ресурсов, содержащихся в выдаче поисковой системы в ответ на формализованный запрос.

Полнота – мера эффективности поиска, выраженная в виде отношения числа релевантных ресурсов, извлечённых поисковой системой из Интернета в ответ на формализованный запрос, к общему количеству релевантных ресурсов, содержащихся в Интернете.

Полнота показывает, насколько хорошо поисковая система находит то, что нужно пользователю; точность показывает, насколько хорошо поисковая система отфильтровывает то, что пользователю не нужно.

В приведённых выше определениях точности и полноты поиска мы можем заменить «релевантный» на «пертинентный». В таком случае получим, что точность – это доля пертинентных ресурсов среди всех ресурсов, присутствующих в выдаче, а полнота – это доля пертинентных ресурсов, присутствующих в выдаче, среди всех пертинентных ресурсов, имеющихся в сети Интернет. Очевидно, что при таком изменении определений полноты и точности, изменятся и их числовые значения для конкретных результатов обработки конкретных запросов. Есть некоторый произвол в определении мер эффективности информационного поиска.

Чтобы сделать наше исследование точнее, потребуется ввести более формальное определение параметров эффективности поиска. Наш подход основан на логическом анализе естественного языка и теории речевых действий [7, 26, 27].

В информационном поиске мы рассматриваем три сущности: человек – пользователь, желающей получить информацию; формализованный запрос; и последняя, третья – это ответ на запрос, представленный в виде упорядоченного списка найденных Интернет-ресурсов.

Поэтому первый шаг поиска информации – в формулировании запроса. Соответствие между информационной потребностью пользователя и формализованным запросом определяет успех информационного поиска. Назовём соответствие между информационной потребностью пользователя и формализованным запросом ***адекватностью*** запроса.

Мы определяем ***релевантность*** как бинарное отношение между формализованным запросом и ответом на этот запрос. Числовое значение релевантности зависит от трех параметров – точности, полноты и ранжирования.

Ранжирование – правильность порядка, в котором представлен список результатов информационного поиска.

Точность – доля релевантных ресурсов среди всех ресурсов, присутствующих в выдаче.

Полнота – это доля релевантных ресурсов присутствующих в выдаче среди всех релевантных ресурсов, имеющихся в сети Интернет.

Пертинентность – это бинарное отношение между информационной потребностью пользователя (формализованной в запросе) и списком Интернет-ресурсов, который поисковая система выдала в ответ на этот запрос; пертинентность зависит от релевантности списка результатов, и адекватности формализованного запроса. Часто пертинентностью называют то, насколько выданный поисковой системой список Интернет-ресурсов интересен пользователю. Мы это определяем иначе: пертинентность – это то, насколько выданный поисковой системой список Интернет-ресурсов соответствует *той информационной потребности, которую пользователь пытался сформулировать в данном формализованном запросе* (средствами, предоставляемыми данной поисковой системой). Различие состоит в следующем.

Ресурсы, представленные в выдаче поисковой системы, могут быть интересны пользователю, но это может быть совсем не то, что он пытался сформулировать в конкретном запросе.

Главная цель нашего исследования – разработка методов повышения ***пертинентности*** информационного поиска. Для получения высокой пертинентности, мы должны достигнуть и высокой

адекватности, и высокой релевантности. Чтобы получать высокую адекватность, пользователь должен иметь: различные и достаточно богатые инструменты создания формализованного запроса, т.е. представления осознаваемой им информационной потребности в формальном виде; возможность делать точную и полную формулировку его информационной потребности.

В терминах наших определений, для пользователя важна пертинентность результата информационного поиска, а не его релевантность. Чем беднее возможности формулировки запроса, тем проще информационно-поисковой системе добиться высокой релевантности. Даже при максимальной релевантности выдачи поисковой системы пертинентность – удовлетворенность пользователя – может быть близкой нулю, если формализованный запрос, сформулированный пользователем, неадекватно отражает его реальную информационную потребность. Такая ситуация возможна из-за неопытности пользователя и сложности требуемой информации. Приведём пример. Пользователь хочет узнать, о чём пишут студенты Новосибирского государственного университета (НГУ). Он вводит в Гугл запрос «форум студентов НГУ». Просмотрев первые 5 страниц выдачи Гугла (т.е. первые 50 выбранных ресурсов), пользователь обнаруживает, что среди них нет ни одной ссылки на какой-либо студенческий форум в НГУ, т.е. пертинентность первых пятидесяти ссылок равна нулю. При этом релевантность, и, в частности, ранжирование выдачи очень высоки. Рассмотрим подробнее утверждение: чем беднее возможности формулировки запроса, тем проще информационно-поисковой системе добиться высокой релевантности, когда в качестве запроса пользователь вводит ровно одно слово. В этом случае поисковая система должна предоставить документы, содержащие как можно большее число вхождений данного слова. Современные поисковые системы для таких запросов покажут очень высокую релевантность. Однако, при высокой релевантности указанного выше запроса, в подавляющем числе случаев удовлетворенность пользователя – пертинентность – будет крайне низкой, а во многих случаях равной нулю (или даже отрицательной – если пользователь потратил своё время и не получил никакой нужной ему информации). Таким образом, для пользователя важно не соответствие формального запроса выдаче, а реальное соответствие выдачи поисковой системы его информационной потребности, т.е. не релевантность, а пертинентность.

Чем беднее язык формулирования запросов и неискушённее пользователь, тем примитивнее получается запрос. А чем примитивнее запрос, тем выше релевантность выдачи и ниже пертинентность.

В результате возможна парадоксальная ситуация, когда удовлетворение информационной потребности пользователя (пертинентность) может быть обратно пропорционально релевантности выдачи поисковой системы.

Потребность пользователя и результат выдачи (список найденных Интернет-ресурсов) можно представить как начало и конец пути, по которому пользователь и поисковая система должны пройти, чтобы доставить пользователю необходимую ему информацию. Момент, когда поисковый запрос сформулирован, находится на этом пути и по существу означает конец работы пользователя и начало работы поисковой системы. Чем примитивнее выразительные возможности формулировки запроса, тем ближе этот момент к концу пути, и тем дальше он от потребностей пользователя. Соответственно, тем большую работу должен проделать пользователь, чтобы решить свою информационную проблему. И наоборот, чем сложнее и выразительнее язык поискового запроса, тем больше «путь» поисковой системы, и тем сложнее ей достигнуть релевантности. В то же время «путь» пользователя меньше, и ему легче сформулировать правильный запрос, чтобы получить нужную информацию.

Рассмотрим инструменты информационного поиска – информационно-поисковые системы и Интернет-каталоги.

Преимущества информационно-поисковых систем, Гугл и Яндекс:

- ✧ высокая релевантность выдачи, т.е. обеспечивают высокое соответствие найденных документов сделанному формальному запросу;
- ✧ используют сходные принципы определения релевантности документов;
- ✧ полнота найденной информации;
- ✧ осуществляют поиск практически по всем ресурсам, представленным в Интернете. Поисковой системой Гугл в настоящее время проиндексировано более 24–х миллиардов страниц;
- ✧ индекс русскоязычной поисковой системы Яндекс содержит более 2–х миллиардов страниц. Поисковые системы постоянно просматривают Интернет. Поэтому пользователь получит подавляющее большинство документов, находящихся в настоящий момент времени в Интернете и релевантных формализованному запросу;
- ✧ точность поиска. Одним из видов поискового шума, т.е. нерелевантных Интернет-ресурсов,

представленных в выдаче, является спам – злонамеренные действия разработчиков вебсайтов, позволяющие веб-странице попадать в поисковые выдачи по запросам, не имеющим никакого отношения к тематике данной страницы. Цель спама – те или иные виды рекламы. Сейчас поисковые системы практически полностью решили проблему борьбы со спамом и достигают очень высокой точности поиска;

- ✧ высокая точность ранжирования. Благодаря ряду разработанных поисковыми системами алгоритмов, таких как, например, определение системой Гугл Пэйдж-ранга Интернет-страницы, найденные ресурсы хорошо упорядочиваются поисковыми системами по степени релевантности формализованному запросу.

Главный недостаток информационно-поисковых систем – проблематичная пертинентность. Проблема в том, что пользователь не всегда может в достаточно полной мере выразить свою информационную потребность, так как обычно он формулирует запрос из нескольких, как правило, двух-трёх слов. Таким способом крайне трудно (порой просто невозможно) сформулировать сложную информационную потребность. Поэтому пертинентность результата работы информационно-поисковой системы – дело случая и везения.

Для оценки работы существующих поисковых систем характеристика пертинентности неприменима, а применима только релевантность. Это означает: давая системе разные запросы, мы можем оценивать релевантность – полноту выдачи, наличие поискового шума, правильность ранжирования найденных Интернет-ресурсов (т.е., их место в списке выдачи) и т.д. Но формально ничего не можем сказать о пертинентности, поскольку запрос из 1–5 слов и даже расширенный запрос могут составить пользователи с совершенно разными информационными потребностями. Например, когда пользователь вводит запрос «теория конструктивных моделей», он хочет найти тексты статей по этой теме, купить учебник по теории конструктивных моделей, найти объявления о конференциях или форум, где можно задать вопросы и т.д. Таким образом, результаты конкретной выдачи для одного пользователя, сформулировавшего данный запрос, могут быть вполне пертинентными, а для другого, написавшего точно такой же запрос, пертинентность может быть нулевой. Но, поисковая система отрабатывает именно данный запрос, поэтому понятие пертинентности выдачи поисковой системы в строгом смысле некорректно. Эта ситуация очень сходна с соотношением между глубинными и поверхностными структурами в лингвистике, которое изучал Н. Хомский.

Таким образом, в результате работы информационно-поисковых систем достигается полнота и актуальность найденной информации. Алгоритмы поиска обеспечивают высокую релевантность и низкое количество поискового шума. Но при этом не решается и корректно не ставится задача достижения пертинентности.

Другой инструмент поиска информации в Интернете – Интернет-каталоги (т.е. каталоги Интернет-ресурсов) – разбитый по рубрикам набор ссылок на Интернет-ресурсы, снабжённых краткими описаниями.

Каталоги имеют ясный и понятный пользователю интерфейс. Все адреса ресурсов упорядочены по темам и организованы в виде древовидной структуры рубрик и подрубрик. Пользователь просматривает рубрикатор и выбирает интересующий его раздел. Каждая рубрика содержит список ссылок на ресурсы Интернета, отвечающих данной тематике.

Поисковые системы и каталоги могут быть специализированными и общего назначения. В специализированных каталогах собраны ссылки на страницы определённой тематики.

Недостатки Интернет-каталогов:

- ✧ маленькое количество ссылок на Интернет–ресурсы;
- ✧ содержат доли процента от всех ресурсов по данной тематике, представленных в Интернете;
- ✧ отсутствие свежей информации. Каталоги нужно постоянно пополнять, но обычно это не делается оперативно. Поэтому каталоги не содержат ссылок на самые новые Интернет-ресурсы;
- ✧ негарантированная релевантность. Составители каталогов, исходя из собственных вкусов, могут помещать в рубрики ресурсы, лишь отчасти соответствующие указанной тематике.

3. ВИРТУАЛЬНЫЙ КАТАЛОГ

В поисковых системах и Интернет-каталогах, мы видим:

- ✧ высокую релевантность отработки запросов поисковыми системами;
- ✧ отсутствие возможности корректного определения пертинентности для поисковых систем;
- ✧ высокую пертинентность каталогов (но только для объёма представленных в них ресурсов).

Если мы хотим добиться релевантности – полноты и точности найденной информации, то пользуясь

поисковыми системами, не достигнем пертинентности, поскольку пользователь не имеет возможности точно и полно сформулировать свою информационную потребность. Если мы хотим предоставить пользователю ясный и удобный интерфейс Интернет-каталога для точного выражения его информационной потребности, то теряем полноту найденной информации.

Предлагаемое нами решение этой проблемы – синтез информационно-поисковых систем и Интернет-каталогов – виртуальный каталог [9], объединяющий преимущества двух представленных выше методов информационного поиска: простоту и ясность каталогов, полноту и актуальность найденной информации, обеспечиваемую информационно-поисковыми системами.

Интерфейс виртуального каталога внешне напоминает интерфейс обычного Интернет-каталога. В его основе система рубрик, соответствующая иерархии подобластей данной предметной области. В качестве названий рубрик каталога берутся разделы и подразделы данной предметной области. Каждой рубрике сопоставлено объяснение её смысла на естественном языке. Наличие описания рубрик устраняет один из недостатков каталогов – отсутствие справочной информации. Пользователь выбирает определённую рубрику и получает список Интернет-ресурсов, которые ей соответствуют.

Однако в отличие от обычного Интернет-каталога, виртуальный каталог не хранит ссылок на конкретные Интернет-ресурсы. Вместо этого по названию рубрики определяется запрос к информационно-поисковой системе. Для обеспечения релевантности информационного поиска каждой рубрике сопоставлен набор специальных эвристик. Эти эвристики – ключевые термины данного раздела предметной области и другие ассоциации к названию рубрики – определяют один или несколько поисковых запросов таким образом, что найденная по нему информация полностью соответствует тематике данной рубрики, следовательно, является той, которую ожидает получить пользователь.

Интерфейс виртуального каталога ясный и понятный пользователю. Навигация по подразделам простая, не отнимает много времени, может осуществляться как в горизонтальном, так и в вертикальном направлениях. Горизонтальное направление – переход между смежными разделами. Вертикальная навигация – от раздела к подразделу, от подраздела к разделу.

Одна только спецификация предметной области не позволяет точно формализовать информационную потребность пользователя. В данной предметной области пользователь может искать различные типы Интернет-ресурсов. Например, если мы рассмотрим поиск научной информации, это может быть сайт электронного журнала; полный текст научной статьи; информация о конференции; сайт научной организации; персональная страница учёного: каталог Интернет-ресурсов по определённой научной тематике; страница Интернет-магазина, в котором продаются научные книги и т.д. Поэтому (в отличие от обычного Интернет-каталога) в виртуальном каталоге пользователю предоставляется возможность указать интересующую его рубрику (т.е. название раздела предметной области) и тип требуемого Интернет-ресурса.

Следующая проблема, которую нужно решить для полной формальной спецификации информационной потребности пользователя, – определение класса задач, которые он хочет решить при поиске информации. При поиске научной информации такими классами решаемых задач могут быть: поиск текста статьи по её выходным данным; поиск точных выходных данных статьи по автору или названию; поиск статей, описывающих данные объекты, свойства или взаимодействия (например, химические реакции для синтеза данного вещества); поиск патентов, относящихся к данному типу устройств; поиск информации о конференциях по данной тематике; поиск ответа (известного специалистам) на определённый научный вопрос и т.д.

Классы решаемых пользователем задач взаимосвязаны с типами Иитернет-ресурсов, которые он хочет найти. Тем не менее, это две разные проблемы. Для их решения потребуются два разных вида онтологий – онтология Интернета и онтология пользователей информационно-поисковой системы. Рассмотрим решаемую пользователем задачу – найти точные выходные данные статьи. Эту задачу можно решить, отыскав:

- ✧ Интернет-ресурс, содержащий полный текст данной статьи;
- ✧ Интернет-ресурс, содержащий полный текст статьи, со ссылкой на данную статью;
- ✧ персональную страницу автора данной статьи;
- ✧ страницу научного отчёта организации, где работает автор данной статьи;
- ✧ страницу издательства журнала, где опубликована данная статья и т.д.

Таким образом, типы Иитернет-ресурсов и классы решаемых пользователем задач взаимосвязаны, тем не менее, они представляют собой разные классификации.

Для наиболее полной и точной формализации информационной потребности пользователю

предоставлена возможность самому указывать дополнительные слова, которые должны (или не должны) присутствовать в требуемых документах.

Инструментами виртуального каталога:

- ✧ достигается *адекватность* формализации информационной потребности пользователя, лежащая в основе пертинентности информационного поиска;
- ✧ обеспечивается полнота, точность и правильное ранжирование найденной информации за счёт использования запросов к поисковым системам. При обработке запроса просматриваются все ресурсы, имеющиеся на данный момент времени в сети Интернет. Это гарантирует высокую релевантность списка найденных Интернет-ресурсов;
- ✧ обеспечивается высокий уровень пертинентности информации, найденной при помощи виртуального каталога.

Таким образом, виртуальный каталог строится на основе трех видов онтологий:

- ✧ иерархия онтологий разделов и подразделов данной предметной области;
- ✧ онтология сети Интернет;
- ✧ онтология пользователя информационно-поисковой системы.

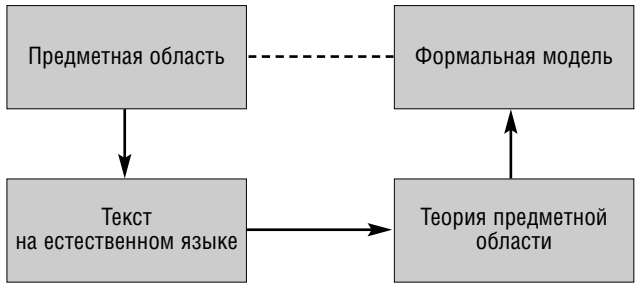
4. ТЕОРЕТИКО-МОДЕЛЬНАЯ ФОРМАЛИЗАЦИЯ ОНТОЛОГИЙ

Онтологии предметных областей – инструмент, необходимый для достижения высокой пертинентности информационного поиска. Дадим точное определение онтологии предметной области в терминах теории моделей и исследуем свойства соответствующих математических структур.

Понятие онтологии предметной области возникло в инженерии знаний. Онтология нужна для описания основных терминов (ключевых понятий) данной предметной области, цель которого – в явном виде определить значение терминов, специфичных для данной предметной области; показывает общее видение таких предметных областей. «Онтология – это явная спецификация концептуализации» [18].

Необходимость использования онтологий вытекает из общей постановки задачи моделирования дискретных систем, представленной на рисунке.

Моделируемая предметная область представлена в виде набора текстов на естественном языке (в большей или меньшей степени структурированных). Задача – построить формальную модель данной предметной области. Для этого сначала нужно построить теорию предметной области.



Первый шаг построения теории предметной области – формальное описание онтологии предметной области, т.е. смысла всех используемых терминов, специфичных для данной предметной области.

Исследованиям по онтологиям посвящены [5, 8, 9, 13–25, 28–30]. Кратко сформулируем, что специалисты подразумевают под понятием онтологии [8]:

- ✧ онтология – инструмент для моделирования реальности;
- ✧ онтология описывает определенную предметную область;
- ✧ знание, представленное онтологией, должно быть интерсубъективным (это означает, что все эксперты в данной предметной области должны признавать утверждения, представленные в онтологии этой предметной области).
- ✧ онтология должна содержать глоссарий ключевых понятий и спецификацию их смысла.

Главная цель онтологии – описывать общие свойства предметной области. Мы рассматриваем онтологию с точки зрения её содержания, определяем онтологию в терминах информации о предметной области. Для формального определения содержания онтологии предметной области применяем подход Р. Карнапа к описанию видов истинности предложений [11, 12]. Он пересмотрел понятие аналитических суждений, введённое И. Кантом, и предложил три типа истинности высказываний: логическая истинность; аналитическая истинность; синтетическая истинность.

Утверждение является логическим (логически истинным или логически ложным) если значение истинности этого утверждения полностью определяется его логической формой. Например, предложение $(\varphi \rightarrow \varphi)$ является логически истинным, а предложение $(\varphi \& \neg \varphi)$ – логически ложным.

Предложение является аналитическим, если его значение истинности зависит только от смысла понятий, содержащихся в этом утверждении. Например, предложение «у холостого мужчины нет жены» аналитически истинное, а предложение «у треугольника четыре угла» – аналитически ложное.

Предложение является синтетическим, если значение истинности этого предложения зависит от реального мира. Например, утверждение «Земля – планета Солнечной системы» синтетически истинное, а утверждение «Земля плоская» – синтетически ложное.

Онтология должна описывать общие свойства предметной области, не зависящие от её конкретной реализации; содержать только ту информацию, которая является верной для каждого примера данной предметной области. Одно из наиболее важных свойств онтологии предметной области – гарантированная возможность её переиспользования, когда мы имеем дело с различными экземплярами данной предметной области. Из этого следует: онтология должна содержать только аналитические предложения [8, 28].

Значение истинности аналитического предложения непосредственно вытекает из смысла понятий, встречающихся в этом предложении. Поэтому оно не зависит от того, какой экземпляр предметной области мы рассматриваем. С другой стороны, для любого утверждения, которое не является аналитическим, можно представить себе ситуацию, где это утверждение будет ложно. Мы можем быть полностью уверены, что предложение является истинным на любом примере данной предметной области, только тогда, когда предложение аналитическое (в том смысле, как понимаются термины данной предметной области).

Таким образом, онтология предметной области должна состоять из набора ключевых понятий предметной области и множества аналитических предложений, дающих полное описание значений этих ключевых понятий.

Определение 1. *Формальной онтологией предметной области SD называется пара* $O=<A, \sigma>$, *где* σ – *множество ключевых понятий предметной области, и A – множество аналитических предложений, описывающих смысл данных ключевых понятий.*

В определении онтологии предметной области множество σ – это сигнатура онтологии. Это означает, что σ содержит только символы понятий. Множество A состоит из определений символов, содержащихся в сигнатуре σ. Кроме того, выполнено $\sigma \subseteq \sigma(A)$, но не обязательно верно $\sigma = \sigma(A)$. Это означает, что множество аналитических предложений A может содержать сигнатурные символы, которые не являются символами ключевых понятий предметной области. Такое может произойти, когда при описании смысла сигнатурных символов (т.е., символов ключевых понятий), мы используем утверждения, содержащие понятия, которые сами не являются ключевыми понятиями данной предметной области.

Для формальной онтологии $O = <A, \sigma>$ множество A не обязательно должно быть теорией, т.е. множество A не обязательно дедуктивно замкнуто.

Наиболее простой вид онтологии, определяющий значения терминов некоторой предметной области, – глоссарий (или тезаурус). Представляет интерес вопрос о возможности представления смысла ключевых понятий произвольной предметной области при помощи глоссария. Для ответа на этот вопрос рассмотрим формальное определение глоссария предметной области.

В современном понимании глоссарий состоит из статей, в которых даётся объяснение (определение) ключевых терминов некоторой предметной области. Статья глоссария состоит из формулирования определения термина и содержательной части, которая более подробно раскрывающей смысл этого термина. Глоссарий описывает определённую область знаний, некоторую предметную область.

В качестве простого примера фрагмента глоссария можно привести следующие «определения»:

«Животное – это ... »

«Собака – это животное, которое ... »

«Болонка – это собака, которая ... » и т.д.

Здесь мы начали с наиболее общего термина – «животное», затем перешли к центральному термину – «собака», а затем стали описывать частные случаи (породы) собаки – болонку и т.п.

Дадим формальное определение глоссария в теоретико-модельных терминах. Для этого нам потребуются некоторые определения и обозначения [3, 4].

Сигнатурой назовем кортеж

$$\sigma = <P_1, \ldots, P_n, f_1, \ldots, f_k, c_m, \ldots, c_m>,$$

где $P_1, \ldots, P_n$ – символы предикатов;

$f_1, \ldots, f_k$ – символы функций (операций);

$c_m, \ldots, c_m$ – символы констант (т.е. выделенных элементов).

Через $S(\sigma)$ обозначим множество всех предложений, т.е. формул без свободных переменных, сигнатуры σ.

Для формулы φ через $\sigma(\varphi)$ обозначим сигнатуру формулы φ, т.е. множество всех сигнатурных символов, входящих в φ. Через $\sigma(\Gamma)$ обозначим сигнатуру множества формул Γ.

Теорией называется дедуктивно замкнутое множество предложений. Это означает, что если

предложение (данной сигнатуры!) выводимо из теории, оно обязательно должно принадлежать этой теории.

Для множества предложений Γ через $Th(\Gamma) = \{\psi \in S(\sigma(\Gamma)) \mid \Gamma \vdash \psi\}$ обозначим теорию, аксиоматизируемую множеством предложений Γ. Через $Th(\varphi) = Th(\{\varphi\})$ обозначим теорию, аксиоматизируемую предложением φ. Символом $\subset$ мы будем обозначать строгое включение. То есть $A \subset B$ означает, что $A \subseteq B$ и $A \neq B$.

Определение 2. Пусть σ – сигнатура. Последовательность предложений $\varphi_1, \ldots, \varphi_n \in S(\sigma)$ назовем *формальным глоссарием (определяющим понятия из* σ), если:

а) $\sigma(\varphi_1) \subset \sigma(\varphi_1 \& \varphi_2) \subset \ldots \subset \sigma(\varphi_1 \& \ldots \& \varphi_n) = \sigma$;

б) добавление каждого нового предложения φ_k консервативно расширяет предыдущий набор предложений, т.е.

$$Th(\varphi_1 \& \ldots \& \varphi_k) = Th(\varphi_1 \& \ldots \& \varphi_n) \cap S(\sigma(\varphi_1 \& \ldots \& \varphi_k)).$$

Консервативность расширения означает следующее: при определении новых понятий мы не должны менять смысл уже определённых понятий.

Мы определили смысл термина только тогда, когда далее в глоссарии его смысл уже не будет переопределяться (в частности, не будет добавляться новая информация, существенная для его смысла). В противном случае, определением термина является весь текст глоссария, т.е. это уже не глоссарий, а одно единое определение набора понятий (терминов). Поэтому консервативность расширения – необходимое условие в определении глоссария.

Определение 3. Будем говорить, что формальный глоссарий $\varphi_1, \ldots, \varphi_n$ представляет множество предложений Γ, если $Th(\Gamma) = Th(\varphi_1 \& \ldots \& \varphi_n)$.

Определение 4. Будем говорить, что формальный глоссарий $\varphi_1, \ldots, \varphi_n$ *явно определяет* понятия из σ, если существуют такие формулы $\psi_1, \ldots, \psi_n$, что для любого $k<n$ выполнено

$$\varphi_{k+1} = \forall \mathbf{x}\, (P(\mathbf{x}) \leftrightarrow \psi_{k+1}(\mathbf{x})), \text{ либо}$$
$$\varphi_{k+1} = \forall \mathbf{x}((f(\mathbf{x}) = y) \leftrightarrow \psi_{k+1}(\mathbf{x}, y)), \text{ либо}$$
$$\varphi_{k+1} = \forall y((c = y) \leftrightarrow \psi_{k+1}(y)),$$

где $P, f, c \in \sigma \setminus \sigma(\varphi_1 \& \ldots \& \varphi_k)$;

$\mathbf{x}$ – кортеж (n-ка) переменных;

$\sigma(\psi_{k+1}) \subseteq \sigma(\varphi_1 \& \ldots \& \varphi_k)$.

Три указанных вида определений – явные определения предиката (n-местного отношения), функции и константы.

Пример фрагмента явного глоссария – приведённая выше последовательность определений: «Животное – это ... », «Собака – это животное, которое ... », «Болонка – это собака, которая ... ».

Всегда ли смысл ключевых понятий предметной области можно задать в виде явного глоссария – последовательности явных определений? Ниже дан отрицательный ответ на этот вопрос. Для этого мы обсудим, всегда ли можно так организовать глоссарий ключевых терминов предметной области, чтобы эти понятия определялись по одному, т.е. одно за другим. В реальных глоссариях, написанных на естественном языке, понятия, как правило, определяются именно так – по одному. Поэтому и возникает вопрос – имеется ли такая возможность в общем случае, т.е., всегда ли возможно такое представление.

Замечание 1. Для произвольного множества предложений S сигнатуры σ существует консервативная последовательность множеств предложений $S_1, \ldots, S_n$ такая, что $Th(S_1 \cup \ldots \cup S_n) = Th(S)$, $\sigma(S_1) \subset \sigma(S_1 \cup S_2) \subset \ldots \subset \sigma(S_1 \cup \ldots \cup S_n) = \sigma$ и для любого $k<n$ множество $\sigma(S_1 \cup \ldots \cup S_{k+1}) \setminus \sigma(S_1 \cup \ldots \cup S_k)$ содержит ровно один символ.

Таким образом, используя бесконечные множества предложений, мы всегда можем построить последовательность определений ключевых понятий предметной области так, чтобы понятия определялись по одному. Однако в реальных глоссариях мы имеем дело только с конечными множествами предложений.

Вопрос. а) Если множество предложений S конечно, можно ли в замечании 1 подобрать последовательность множеств предложений $S_1, \ldots, S_n$ так, чтобы все теории $Th(S_k)$ были конечно аксиоматизируемыми?

б) Верно ли, что для произвольного предложения φ существует формальный глоссарий $\varphi_1, \ldots, \varphi_n$ такой, что $Th(\varphi) = Th(\varphi_1 \& \ldots \& \varphi_n)$ и для любого $k<n$ множество $\sigma(\varphi_1 \& \ldots \& \varphi_{k+1}) \setminus \sigma(\varphi_1 \& \ldots \& \varphi_k)$ содержит ровно один сигнатурный символ?

Следующая теорема даёт отрицательный ответ на оба пункта этого вопроса.

Теорема. Существует сигнатура $\sigma = \{s_1, s_2\}$ и предложение φ, определяющее понятия из σ, для которого нет формального глоссария φ_1, φ_2, представляющего $\{\varphi\}$, такого, чтобы $\sigma(\varphi_1) \subset \sigma(\varphi_2)$ (т.е., $\sigma(\varphi_1) \subset \sigma(\varphi_2)$) и $\sigma(\varphi_1) \neq \varnothing$.

Из теоремы непосредственно вытекают:

- **Следствие 1.** В общем случае онтология не может быть представлена в виде глоссария, определяющего понятия одно за другим.

Вернёмся теперь к вопросу – может ли произвольная онтология быть представлена явным глоссарием? Всегда ли смысл набора понятий может быть представлен явным глоссарием? Явный глоссарий даёт определения понятий одно за другим – т.е., по одному. Поэтому мы получаем отрицательный ответ и на этот вопрос;

- ✧ **Следствие 2**. В общем случае смысл ключевых понятий предметной области не может быть представлен в виде явного глоссария.

5. ЗАКЛЮЧЕНИЕ

Для решения задачи точного поиска информации нами соединены два подхода: методы поисковых систем для обеспечения релевантной отработки формального запроса; интерфейс Интернет-каталогов, позволяющий обеспечить пользователю понятный и удобный интерфейс. Мы решаем задачу поиска информации в сети Интернет с помощью виртуального каталога. Система рубрик виртуального каталога основана на иерархии онтологий предметных областей; при помощи этой иерархии онтологий достигается релевантность найденныхдокументов выбранной предметной области.

Пертинентность информационного поиска достигается за счёт спецификации не только предметной области, в которой ищется информация, но и вида требуемого Интернет-ресурса, а также типа поисковой задачи, которую хочет решить пользователь. Для этого используются онтология сети Интернет и онтология пользователя информационно-поисковых систем.

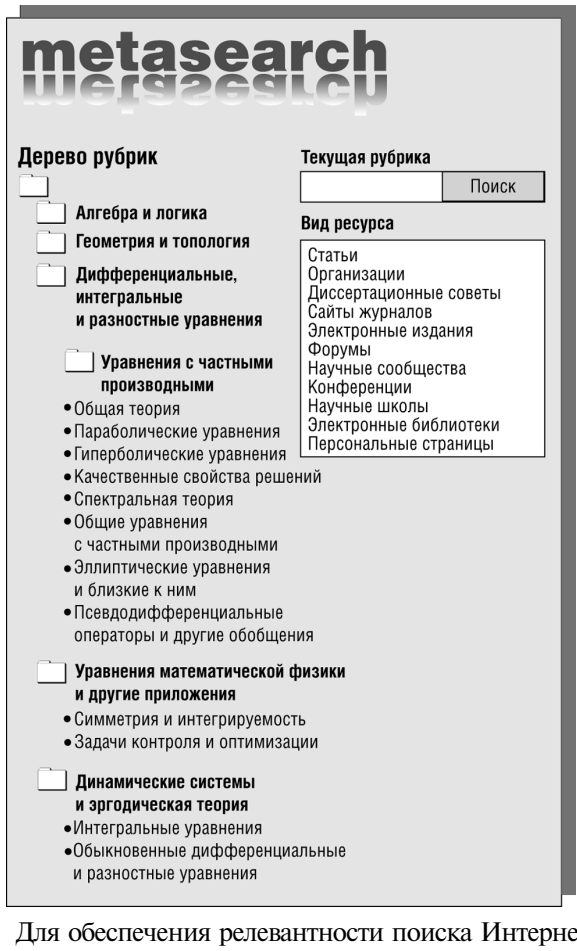
Разработка онтологий предметных областей ведётся на основе теоретико-модельного подхода к формализации онтологий. Онтология предметной области представляется в виде пары – сигнатуры предметной области, состоящей из ключевых терминов, и множества аналитических предложений, описывающий смысл ключевых терминов предметной области.

Предложена теоретико-модельная формализация глоссария предметной области. Показано, что не всегда явного глоссария достаточно для спецификации смысла ключевых терминов данной предметной области.

Технологии, разрабатываемые в рамках данного подхода, применяются для создания метапоисковых систем для поиска в Интернете научно-технической информации по математике, химии (катализу) и патентоведению. Эти системы – реализация идеи виртуального каталога; их интерфейс – иерархия рубрик по каждой из указанных предметных областей.

При работе с такими системами пользователь находит интересующую его рубрику и запускает поиск. После этого ищутся Интернет-ресурсы, наиболее релевантные выбранной рубрике. Пользователь может выбирать рубрику любой степени вложенности, например, «Алгебра и логика», «Логика», «Теория вычислимости» и т.д. Найденные Интернет-ресурсы будут соответствовать именно выбранной рубрике, независимо от глубины её вложенности.

Кроме указания рубрики, пользователь может указать вид требуемого Интернет-ресурса. Это может быть: «Статьи», «Организации», «Сайты журналов», «Электронные издания», «Форумы», «Научные сообщества», «Конференции», «Электронные библиотеки», «Персональные страницы» и др. Таким образом, достигается более точная формулировка пользователем его запроса.



Для обеспечения релевантности поиска Интернет-ресурсов, соответствующих выбранной рубрике и имеющих указанный тип, используются специальные эвристики. Методы порождения таких эвристик подробно рассмотрены в нашей работе, по технологиям практической реализации виртуальных каталогов. ■

Литература

1. Гультяев А.К. Поиск в Интернете. 2-е издание. Питер, 2006.
2. Гусев В.С. Google – эффективный поиск. Диалектика, 2006.
3. Ершов Ю.Л., Палютин Е.А. Математическая логика. Москва, Наука, 1979.
4. Кейслер Г., Чэн Ч.Ч. Теория моделей. Москва, Мир, 1977.
5. Клещев А.С., Артемьева И.Л. Математические модели онтологий предметных областей. Части 1–3. Научно-техническая информация, серия 2 «Информационные процессы и системы», 2001, № 2, С. 20–27, № 3, С.19–29, № 4, с. 10–15.
6. Ландэ Д.В. Поиск знаний в Internet. Издательский дом «Диалектика-Вильямс».
7. Пальчунов Д.Е. Алгебраическое описание смысла высказываний естественного языка. Модели когнитивных процессов. Новосибирск, 1997 – Вып. 158: Вычислительные системы, стр. 127–148.
8. Пальчунов Д.Е. Моделирование мышления и формализация рефлексии I: Теоретико-модельная формализация онтологии и рефлексии. Философия науки, № 4(31), 2006, с.86–14.
9. Пальчунов Д.Е., Сидорова Е.С. Виртуальный каталог. Труды Всероссийской конференции «Знания–Онтологии–Теории», Новосибирск, 2007, стр. 166–175.
10. Холмогоров В. Поиск в Интернете и сервисы Яндекс. Питер, 2006. ГОСТ 7.73-96
11. Carnap, R. Meaning and Necessity. A Study in Semantics and Modal Logic. Chicago, 1956.
12. Carnap, R. Philosophical Foundations of Physics. Basic Books, New York, London, 1968.
13. Daconta M.C., Obrst L.J., Smith K.T. The Semantic Web: A Guide to the Future of XML, Web Services, and Knowledge Management. Wiley Publishing, 2003.
14. Fensel D. OIL: An Ontology Infrastructure for the Semantic Web. IEEE Intelligent Systems 16, 2, 2001.
15. Fensel D. Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce. Springer Verlag, 2004.
16. Gangemi A., Pisanelli D. M., Steve G. An Overview on the ONIONS Project: Applying Ontologies to the Integration of Medical Terminologies. In: Data & Knowledge Engineering, Vol. 31, N 2, 1999,183–220.
17. Gomez-Perez A. Ontology Engineering. Springer Verlag, 2002/2003.
18. Gruber, T. R. A Translation Approach to Portable Ontologies. Knowledge Acquisition, 5(2), 1993, 199–220.
19. Gruber, T. R./Olsen, G. R. An Ontology for Engineering Mathematics. In: Doyle, Jon/ Torasso, Piero/Sandewall, Erik (Eds.): Fourth International Conference on Principles of Knowledge Representation and Reasoning, Gustav Stresemann Institut, Bonn, Germany, Morgan Kaufmann, 1994.
20. Gruber T. R. Towards Principles for the Design of Ontologies Used for Knowledge Sharing. International Journal of Human-Computer Studies Vol.43, Issue 5–6, Nov./Dec. 1995, 907– 928.
21. Guarino N. Formal Ontology and Information Systems. In: N.Guarino (ed.) Proceedings of International Conference on Formal Ontology in Information Systems (FOIS'98), Trento, Italy. Amsterdam, IOS Press, 1998, 3–15.
22. Inaba, A./Mizoguchi, R. Learning Design Palette: An Ontology-aware Authoring System for Learning Design. Proc. of International Conference on Computers in Education, (ICCE2004), Melbourne, Australia, Nov. 30 – Dec. 3.
23. Maedche A. Ontology Learning for the Semantic Web. Kluwer Academic Publishers, 2002.
24. McGuinness D., Harmelen F. (eds.) OWL Web Ontology Language Overview. 2003.
25. Mizoguchi R. Ontological Engineering: Foundation of the next generation knowledge processing, N.Zhong et al. (Eds.) WI2001, LNAI2198, Springer-Verlag, 2001, 44–57.
26. Pal'chunov, D. E. Algebraische Beschreibung der Bedeutung von ?u?erungen der nat?rlichen Sprache. In: Zelger, Josef/Maier, Martin (Hrsg.): GABEK. Verarbeitung und Darstellung von Wissen. Innsbruck–Wien: STUDIENVerlag, 1999, 310–326.
27. Pal'chunov D. E. On a logical analysis of GABEK. In: Buber, Renate/Zelger, Josef (Hrsg.): GABEK II. Zur Qualitativen Forschung On Qualitative Research. Innsbruck–Wien–Munchen: STUDIENVerlag, 2000, 185–203.
28. Pal'chunov D. E. GABEK for Ontology Generation. In: Herdina P., Oberprantacher A., Zelger, J. (eds.): Learning and Development in Organizations. (GABEK – Contributions to Knowledge Organization, Vol. 2), Wien: LIT, 2007, p. 87–107.
29. Staab S., Studer R. (eds.) The Handbook on Ontologies in Information Systems. Springer Verlag, 2003.
30. Wielinga B. J., Schreiber A. Th. Reusable and Sharable Knowledge Bases: A European Perspective. In: K. Fuchi, (ed.); Proceedings KB&KS'93, International Conference on Building and Sharing of Very Large-Scale Knowledge Bases'93, JIPDEC, Tokyo, 1993, 103–115.