Поиск в научной электронной библиотеке на основе логико-семантической сети «Вопрос – ответ – реакция»

© В.Н. Добрынин[1], И.А. Филозова[2]

[1] Международный университет природы, общества и человека «Дубна»

[2] Объединенный институт ядерных исследований, г. Дубна

arbatsolo@yandex.ru, Irina.Filozova@jinr.ru

Аннотация

В статье излагается описание технологии семантического поиска в электронных библиотеках на основе логико-семантической сети «Вопрос – ответ – реакция» (ЛСС ВОР), адекватной смыслу текста. Такая поисковая система позволит пользователю получить ответ на вопрос, сформулированный на естественном языке. Излагаются имеющийся опыт и наработки в данной области. Приводится краткий обзор известных вопросно-ответных поисковых систем. Обсуждается возможность применения такого подхода для поиска и навигации в электронном архиве Объединенного института ядерных исследований (ОИЯИ).

1 Введение

Информационные потребности пользователей научных электронных библиотек (ЭБ) определяются задачами, которые они решают в процессе своей профессиональной деятельности. Этими задачами могут быть: исследование, экспертиза, инженерная задача, конструкторская задача, научное руководство и пр. Коммуникация специалистов-профессионалов в данной предметной области эффективна, когда она происходит по принципу вопрос – ответ. Когда пользователь решает некоторую задачу, то, обращаясь к информационной системе, он хочет получить ответ на свой вопрос. Но, приступая к решению новой для себя задачи, пользователь может неточно и/или неполно сформулировать свой вопрос. Это естественно, т. к. полная ясность наступает, когда задача уже решена. Или пользователь, задающий вопрос, не является профессионалом в данной предметной области. Следствие – другой ответ. Тогда возникает типичная проблема: ответ есть в системе, но он не найден, т. к. вопрос сформулирован неточно. Зачастую пользователю трудно понять, является ли полученный им ответ релевантным поставленному им вопросу. Но если пользователь сможет уточнять свой запрос в процессе поиска, он получит новую формулировку вопроса, что приблизит его к осознанию исследуемой проблемы.

Работа специалиста-профессионала с фондами предполагает наличие системы каталогизации и классификации материала. В рамках заданной проблемной темы предмета предлагается технология формирования и поддержки «каталожной» службы, которая обеспечивает эффективный поиск ответов на вопросы. Стержнем такой «каталожной» службы является упорядоченное открытое множество множеств ЛСС ВОР. Множество исходных документов фонда представляются как корпус, ориентированный не только на пользовательские вопросно-ответные потребности, но и на возможность его использования для решения лингвистических задач, связанных с языковыми особенностями документов фонда. Результаты решения лингвистических задач могут служить базой для семантической дифференциации ЛСС ВОР.

2 Семантические поисковые системы

Стандартные поисковые системы выдают список ссылок на найденные ресурсы. Навигацией в массиве найденных ссылок, анализом страниц и поиском необходимой информации пользователь вынужден заниматься самостоятельно.

В настоящее время семантические поисковые системы активно создаются, развиваются и совершенствуются. Они дают лучший результат, чем традиционные поисковые системы, т. к. понижается уровень информационного шума за счет исключения спама и рекламы, другой лишней информации. Но, тем не менее, это не ответ на вопрос пользователя, а список ресурсов, где он может найти ответ. Поэтому поиск ответов на вопросы пользователей, заданных на естественным языке, – актуальная задача.

2.1 Краткий обзор информационно-поисковых систем «Вопрос – ответ»

Вопросно-поисковая система (QA-система) – это информационно-поисковая интеллектуальная справочная система с естественно языковым интерфей-

Труды 12[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2010, Казань, Россия, 2010

сом. Первые QA-системы появились в 1960-е гг. и использовались как естественно-языковые оболочки для экспертных систем.

Среди современных национальных разработок следует упомянуть – Nigma (http://www.nigma.ru), AskNet (http://www.asknet.ru), Генон (http://www.genon.ru/), среди зарубежных – Lexxe (http://www.lexxe.com/), Start (http://start.csail.mit.edu/), Hakia (http://hakia.com), Powerset (http://www.powerset.com). Ниже приводится краткое описание перечисленных систем.

Интеллектуальная поисковая система Nigma.ru – – первая кластеризующая поисковая система в Рунете. Цель проекта – создание программного обеспечения, позволяющего анализировать проиндексированные документы и выдавать конкретную информацию на запрос пользователя, а не ссылки на другие сайты. Созданы такие сервисы, как Nigma-математика, Nigma-химия и Nigma-музыка. Разработки ведутся при участии Стэнфордского университета.

Другой пример – система AskNet, которая обеспечивает поиск ответов на запросы пользователей на русском и английском языках. В качестве результата поиска непосредственно выдаются ответы. Поисковая машина AskNet Global Search осуществляет поиск информации с использованием интернет-сервисов существующих поисковых систем и интернет-энциклопедий по запросам удаленных пользователей.

Система Генон является симбиозом вопросно-ответных и поисковых систем. В Геноне реализована модель накопления и хранения проверенной информации вместе с вопросами, на которые эта информация отвечает. Добавлять свои знания может каждый человек. Ответы и вопросы создаются Авторами, зарегистрировавшимися на Геноне, и проверяются Редакторами на предмет соответствия правилам написания вопросов и ответов (http://www.genon.ru/rules.aspx). Таким образом, базу Генона формируют вопросы, на которые есть однозначные, полные и актуальные ответы, не содержащие какого-либо информационного мусора и прямой рекламы товаров или услуг.

Поисковая машина Lexxe использует полностью автоматическую технологию поиска без участия редактирования пользователем. Большинство ответов приходит из неструктурированных текстов и веб-страниц в интернете. В Lexxe применяется вычислительная лингвистика, что позволяет получить более релевантные результаты, чем от обычных поисковых систем. Lexxe достигает этой цели путем анализа и извлечения значения из поискового запроса. Lexxe распознает, когда запрос является вопросом, а затем пытается найти ответ в Сети, извлекая потенциальные ответы с веб-страниц. Если запрос не является вопросом, производится поиск по ключевым словам.

Поисковая система Start была разработана группой InfoLab в Массачусетском технологическом институте информатики и лаборатории искусственного интеллекта в 1993 г., является универсальной системой. В настоящее время может ответить на миллионы вопросов на английском языке об объектах (города, страны, водоемы, координаты, погода, карты), фильмах (названия, актеры, режиссеры), персоналиях (даты рождения, биографии), терминах и др.

Поисковая машина Hakia производит поиск во всех сегментах, включая веб-новости, блоги, галереи. Новости, блоги, галереи обрабатываются на основе собственной технологии, называемой QDEXing. Веб, видео и изображения обрабатываются Hakia с помощью алгоритма SemanticRank.

Проведение глубокого исследования этих систем затруднено из-за того, что не для всех перечисленных систем представлена в открытом доступе необходимая для полномасштабного исследования информация. Поэтому был проведен экспресс-анализ в виде следующего эксперимента – в системы вводился один и тот же вопрос, на который заведомо известен однозначный правильный ответ. В данном случае это был вопрос «*Где был открыт 105-й элемент периодической таблицы?*». В англоязычные системы вводился тот же вопрос на английском языке. Оценивались показатели: общее количество полученных ответов на первой итерации поиска (∑), количество ответов после фильтрации при наличии в системе соответствующих фильтров (∑F), количество релевантных ответов (∑R). Результаты представлены в таблице, из которой видно, что наиболее высокие показатели наблюдаются у AskNet:

Система	***∑***	***∑F***	***∑R***	***Примечание***
Powerset	6	—	0	
Lexxe	100	—	0	
Start	0	—	0	
Hakia	0	—	0	
AskNet	24	—	7	
Nigma	25 000	9800	7	Из первых 100
Генон	10	—	1	

Качество поиска в описанных системах, несомненно, выше, чем в традиционных поисковых машинах. Эти системы являются универсальными и позволяют задавать любые вопросы по всем областям знаний. В основе их работы лежат лингвистические механизмы – морфологический и синтаксический анализ. Все вопросно-ответные машины ориентированы на получение релевантных ответов на вопросы в широких тематических предметных полях. Теоретические основы таких машин имеют общие и частные подходы. Но они не могут служить основой для создания «каталожной» основы корпуса научных документов.

Подход, излагаемый в данной статье, предполагает создание социотехнической системы *{Эксперт + Инструмент}*. Основная идея – создать дополнительный инструментарий для специализированного фонда документов, содержащий научные тексты, протоколы, проекты и т. д.

3 Семантическая поисковая система на основе логико-семантической сети «Вопрос – ответ – реакция»

Рассматриваемый в статье подход является экспериментальным и основан на качественном анализе научных текстов.

Информационный поиск документов по запросу всегда подразумевает определенную степень осознания пользователем: прежде чем составить информационный запрос, пользователь либо осознанно представляет себе, на какой вопрос хочет получить ответ, либо не может сформулировать вопрос и представляет себе, какая ему необходима информация, чтобы удовлетворить информационную потребность. В последней ситуации он реализует поиск методом проб и ошибок, в процессе которого может либо ясно осознать, на какой вопрос искал ответ, либо не получить желаемого. Для профессионалов всегда есть осознание той информации, которая является ответом на его вопрос. В этом случае ему проще сформировать вопрос и иметь возможность при получении ответа с помощью специализированного навигатора либо уточнять вопрос, либо его углублять, получая соответствующие связки (вопрос – ответ). Тем самым пользователь может получить расширенные, углублённые, уточнённые или новые знания. При этом за счёт реакции пользователь может контролировать согласованность смыслового собственного понимания вопросов и ответов и понимания вопросов и ответов, заложенных в семантической поисковой системе. Поскольку система открытая, пользователь в процессе взаимодействия может уточнять и расширять саму ЛСС.

Общие положения, которые легли в основу данной работы, таковы:

- структурированная, слабо структурированная информация предметной области может быть представлена в виде логико-семантических сетей «Вопрос – ответ – реакция»;
- логико-семантическая сеть отражает определённую тему предметной области;
- предметная область представляется множеством тем;
- предметная задача может быть представлена в форме вопроса (или совокупности вопросов);
- решение задачи может быть представлено в форме ответа или совокупности ответов;
- способ решения задачи может быть представлен унифицированным механизмом поиска реакций на множестве логико-семантических сетей;
- качество решения задачи представляется как уровень релевантности ответов заданному вопросу.

Предполагается, что:

- логико-семантические сети «Вопрос – ответ – реакция» являются основой для структурирования произвольных текстов научно-технической информации;
- ЛСС являются основой структурирования знаний предметной области;
- поиск релевантной информации по запросу может осуществляться на основе унифицированного механизма поиска на ЛСС.

Таким образом, создание семантической поисковой системы на основе ЛСС ВОР включает как стадии разработки:

- теоретических положений технологии поиска ответов на вопросы для специализированных научных корпусов;
- автоматизированной технологии формирования и поддержки специализированных научных корпусов;
- структурно-функциональной модели семантической-поисковой системы на основе ЛСС ВОР;
- прототипов компонент системы;
- технического задания.

3.1 Общие положения системы «Вопрос – ответ»

Вопросы возникают там, где есть познавательная неопределенность. Вопрос не является суждением, но в основе его всегда лежат суждение или совокупность суждений. Раздел логики, изучающий вопросы (эротетическая логика), рассматривает логическую связку «вопрос – ответ» как «единицу мысли».

Развитие научных и практических знаний протекает как переход от ранее установленных суждений к новым, более точным и богатым по содержанию и представляет собой последовательность этапов: постановка вопроса; поиск новой информации; формирование ответа на поставленный вопрос.

Познавательная функция вопроса связана с восполнением, уточнением и конкретизацией ранее полученных общих представлений о предметах и явлениях действительности. В процессе познания вопросы не возникают сами по себе. Любой вопрос всегда опирается на уже известное знание, выступающее его базисом и выполняющее роль предпосылки вопроса (datum questionis).

Познавательная функция вопроса реализуется в форме ответа на поставленный вопрос. Ответ представляет собой новое суждение, уточняющее или дополняющее прежнее знание в соответствии с поставленным вопросом. Поиск ответа предполагает обращение к конкретной области теоретических или эмпирических знаний, которую называют *областью поиска ответа*. Полученное в ответе знание может служить базисом для постановки новых, более глубоких вопросов о предмете исследования. Постановка вопроса и поиск информации для конструирования ответа составляют вопросно-ответную логическую форму развития знаний.

Формулирование вопроса связано с его познавательной функцией, направленной на получение уточняющей информации, или расширенного знания, или углубленного знания, или нового знания об объекте исследования. Вопрос, представленный в форме текста, включает ключевые слова и словосо-

четания, относящиеся к предмету исследования. Между ключевыми словами и словосочетаниями зафиксированы (формой предложения вопроса) определённые отношения. Множество ключевых слов вопроса и отношений между ними образует так называемую онтологическую модель вопроса (ОМВ). Процесс формирования вопроса должен опираться на внутреннюю логику его организации, которая отражается в ОМВ. Это обстоятельство является основой для технологии формирования вопроса.

Процесс поиска ответа на вопрос направлен на поиск информации в области предмета исследования или расширенной области знаний об объекте исследования. Ответ, представленный в форме текста, содержит ключевые слова и отношения между ними. Множество ключевых слов и отношений образуют онтологическую модель ответа. Онтологическая модель ответа и система правил, устанавливающая целостность системы «вопрос – ответ» – основа для технологии поиска ответа на вопрос.

Процесс установления связи вопроса и ответа направлен на выявления возможных несоответствий в ответе. В этом случае необходимо расширить либо область поиска ответа, либо область предпосылки вопроса или предмета исследования. Это должно привести к формированию вопроса, либо его уточнению, либо расширению.

Любой возможный ответ на тот или иной вопрос должен удовлетворять стандартным «постулатам Хэмблина» [1]:

- ответ на вопрос должен быть сформулирован в виде высказывания или предложения;
- возможные ответы на вопрос образуют исчерпывающее множество взаимно исключающих альтернатив;
- значение вопроса известно, если и только если известно, что может считаться ответом на этот вопрос; иными словами, сущность вопроса можно отождествить с множеством возможных ответов на него [3].

По содержанию и структуре ответ должен строиться в соответствии с поставленным вопросом. Лишь в этом случае ответ расценивается как релевантный, т.е. как ответ по существу поставленного вопроса, выполняющий свое основное назначение – уточнить неясную или неопределенную и доставить новую информацию.

Если в качестве ответа приводят хотя и истинные, но содержательно не связанные с вопросом суждения, то их расценивают как ответы не по существу вопроса и обычно исключают из рассмотрения. Появление таких ответов в дискуссии – либо результат заблуждения, когда отвечающий не уловил смысл вопроса, но пытается отвечать на него, либо сознательное стремление уйти от невыгодного ответа на поставленный вопрос.

Логическая зависимость между вопросом и ответом означает, что качество ответа во многом определяется качеством вопроса. На расплывчатый и двусмысленный вопрос трудно получить ясный ответ. Для получения точного и определенного ответа необходимо сформулировать точный и определенный вопрос.

Под точностью и определенностью в данном случае имеется в виду логическая, т. е. понятийно-структурная характеристика вопроса. Она выражается в точности употребляемых понятий и вопросительных слов, а также в рациональном использовании сложных вопросов. Двусмысленные понятия нередко используются в улавливающих или «провокационных» вопросах, в которых содержится скрытая информация. Неопределенность в ответах может быть результатом неясности используемых при постановке вопроса понятий.

Итак,

- качество ответа зависит от качества вопроса;
- под качеством вопроса и ответа следует понимать точность и определённость;
- под точностью и определенностью имеется в виду понятийно-структурная характеристика вопроса и ответа;
- понятийно-структурная характеристика вопроса выражается в точности употребляемых понятий и вопросительных слов, а также в рациональном использовании сложных вопросов;
- понятийно-структурная характеристика ответа выражается в точности употребляемых понятий, а также в рациональном использовании сложных ответов.

3.2 Формальная структура вопроса и ответа

Будем считать, что логическая структура вопроса включает следующие составляющие: тему вопроса (ТВП); содержание вопроса (СВП); объём вопроса (ОВП).

Под темой вопроса будем понимать предпосылки вопроса (знания об объекте и предмете исследования, к которым относится вопрос).

Под содержанием вопроса будем понимать предметы, явления, процессы, технологии, инструменты, теории, относящиеся к объекту и предмету исследования, связи между ними посредством фиксации общих и специфических признаков. Будем считать, что содержание вопроса отражается в ключевых терминах и отношениях между ними, т. е. в онтологической модели вопроса.

Под объёмом вопроса будем понимать множество адекватных ответов, удовлетворяющие заданному уровню качества.

Таким образом, структура вопроса (СТВП) представляется как СТВП =СТВП(ТВП, СВП, ОВП).

Будем считать, что логическая структура ответа включает следующие составляющие: тему ответа (ТОТ); содержание ответа (СОТ); объём ответа (ООТ).

Под темой ответа будем понимать область поиска (знания об объекте и предмете исследования, к которым относится вопрос).

Под содержанием ответа будем понимать предметы, явления, процессы, технологии, инструменты, теории, относящиеся к объекту и предмету исследо-

вания, связи между ними посредством фиксации общих и специфических признаков. Будем считать, что содержание ответа отражается в ключевых терминах и отношениях между ними, т. е. в онтологической модели ответа.

Под объёмом ответа будем понимать множество адекватных вопросу ответов, удовлетворяющее заданному уровню качества. Таким образом, структура ответа (СТОТ) представляется как СТОТ=СТОТ(ТОТ, СОТ, ООТ).

3.3 Формальная связь вопроса и ответа

Будем считать, что вопрос и ответ образуют целостную систему, если удовлетворяются следующие условия:

A. Тема вопроса совпадает с темой ответа, т. е. ТВП=ТОТ (знак «=» означает совпадение, согласованность, адекватность тем);
B. Содержание ответа не больше содержания вопроса (число ключевых терминов в вопросе не меньше числа ключевых терминов в ответе и пересечение множества терминов вопроса и множества терминов ответа не пустое);
C. Объём вопроса не меньше объёма ответа (множество ответов вопроса на предпосылках вопроса больше чем множество ответов области поиска).

Из сказанного следуют ситуации:

- если содержание и объем вопроса совпадает с содержанием и объемом ответа, то ответ и вопрос удовлетворяют качеству и образуют полную систему;
- если содержание ответа является частью содержания вопроса и объём ответа соответствует части объёма вопроса, то ответ частично и качественно соответствует вопросу, и они образуют неполную систему;
- если содержание ответа не соответствует содержанию вопроса, то вопрос и ответ не образуют системы и не удовлетворяют качеству.

3.4 Методика анализа научных текстов

Данная методика применима только к научным текстам. Документ исследуется экспертом с точки зрения:

1. смыслового соответствия названия и содержания;
2. набора фильтров:
   - F1 – общая часть; включает анализ проблемы, ее историю, обзор, актуальность;
   - F2 – авторские понятия; включает вводимые авторами новые термины, общеупотребительные термины с авторской интерпретацией, сужающие семантику;
   - F3 – примеры и иллюстрации; предназначен для пояснения сложных мест в тексте, позволяет сократить размер текста при строгих ограничениях по объему;
   - F4 – идея автора; описывает и раскрывает основную авторскую идею;
3. формирования базовых вопросов, на которые отвечает текст.

На полученном таким образом материале далее строится ЛСС нижнего уровня:

1. Из названия научного текста извлекаются ключевые слова; формируется тезаурус;
2. Выдвигаются гипотезы, о чем идет речь в тексте;
3. Текст разбивается на несколько частей (информационных блоков), к которым применяются фильтры F1, F2, F3, F4;
4. Для частей текста, которые попали в фильтр F4, формулируется основная мысль – несколько предложений; таким образом, мы получаем сжатое изложение (выжимку) текста.

Сопоставив полученную выжимку с выдвинутыми гипотезами, мы получаем возможность делать выводы о том, насколько название текста соответствует его реальному содержанию. После этого этапа можно приступать к формированию ЛСС нижнего уровня: формулирование вопросов к выделенным информационным блокам; выделение ответов из анализируемого текста и ссылок на них; формирование реакций вопросов и ответов; для ЛСС научного текста (нижнего уровня) реакции вопросов и ответов формируются из информационных блоков по фильтру F1.Общая часть, а также по библиографическим ссылкам; формирование графа вопрос – ответ – реакция.

3.5 Логико-семантическая сеть «Вопрос – ответ – реакция»

Прототип семантической поисковой системы на основе ЛСС описан в работе [2]. Логико-семантическая сеть – это множество вопросов, ответов и связей между ними, образующее целостную систему [1]. Целостность ЛСС определяется следующими свойствами:

- ✓ множество «Вопрос – ответ – реакция» относится к определённой теме предметной области;
- ✓ это множество иерархически упорядочено по принципу «от общего к частному»;
- ✓ на нечётном уровне иерархии расположены вопросы, на чётном уровне – ответы и реакции;
- ✓ вопросы *i*-го уровня иерархии связаны только и только с ответами *i*+1-го уровня;
- ✓ вопросы *i*+1-го уровня могут быть связаны с ответами *i*-го уровня;
- ✓ вопрос *i*-го уровня семантически связан с ответами *i*+1-го уровня, если удовлетворяет определенному условию ‘A’ или ‘B’. В случае удовлетворения условию ‘A’, например, имеет место конечная вершина, а в случае удовлетворения условию ‘B’ из данного ответа следуют вопросы *i*+2-го уровня;
- ✓ на *i*=1-м уровне находятся вопросы, которые раскрываются множеством ответов *i*=2-го уровня, частично или полностью охватывающим тему предметной области;
- ✓ на *i*=3-м уровне находятся вопросы, которые дополняют и уточняют ответы *i*=2-го уровня.

Таким образом, ЛСС ВОР можно представить в виде графа (рис. 1).

Вопрос – это выраженный в форме вопросительного предложения запрос, направленный на развитие – уточнение или дополнение знаний.

Ответ – это реализация познавательной функции вопроса в форме вновь полученного суждения. При этом по содержанию и структуре ответ должен строиться в соответствии с поставленным вопросом. Лишь в этом случае ответ расценивается как релевантный, т. е. как ответ по существу поставленного вопроса.

Реакция – это смысловое описание вопроса и ответа [1].

Ввод реакций помогает пользователю понять, получил ли он релевантный ответ на свой вопрос. В качестве реакций могут выступать дополнительная информация по теме вопроса и ответа, ссылки на сайты, словари, рубрикаторы, каталоги и т. д. Такими реакциями могут сопровождаться как вопрос, так и ответ, что позволит пользователю лучше и быстрее сориентироваться в предметной области.

Типы реакций:

- реакции вопроса – это описание предобласти вопроса (для осознания обстоятельств и причин возникновения вопроса и дальнейшего установления смыслового соответствия с областью ответа);
- реакции ответа – это описание области ответа (для осознания смысла вопроса и смысловой связи с ответом).

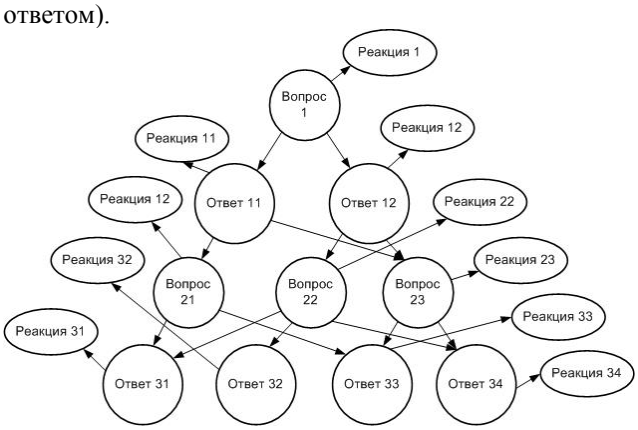


Рис. 1. Граф ЛСС «Вопрос – ответ – реакция»

Процесс постановки вопроса, поиска ответа на него и формирования реакций вопроса и ответа – сложный многоэтапный процесс, описанный подробно в работе [4].

В целом ЛСС ВОР полностью отвечает постулатам Хэмблина.

Реакция для вопроса – это описание области предпосылки вопроса. Реакция ответа – это описание области поиска ответа

Описанная выше методика была применена для построения ЛСС научной статьи: Белага В.В., Семчуков П.Д., Стеценко М.С. Разработка программной оболочки для мультимедийного образовательного продукта // Электронный журнал «Системный анализ в науке и образовании», Международный университет природы, общества и человека «Дубна», кафедра САУ. – 2009, Вып. 2. –http://www.sanse.ru/archive/11. Здесь не представляется возможным дать полное описание процесса вместе с разметкой текста указанной статьи, поэтому в Приложении приведен только полученный граф ЛСС. Отражена сквозная нумерация вопросов и ответов, выполненная экспертом, работавшим с текстом. По свойствам ЛСС она может быть преобразована к виду (ij), где i – номер уровня, j – номер вопроса/ответа данного уровня.

3.6 Формальное представление предметной области

Любая научно-практическая область знаний включает предмет исследования, который может быть представлен проблемным полем (перечнем проблемных вопросов), являющийся основой для научной и практической деятельности. Проблемные вопросы могут быть представлены в виде иерархического дерева по принципу «от общего к частному». Для некоторых вопросов уже существуют возможные альтернативные ответы и способы их реализаций (реакции). Для понимания вопроса также необходима определённая реакция. В свою очередь ответы могут порождать вопросы. Таким образом, проблемный вопрос соотносится к определённой теме предметной области и раскрывается семантической структурой вопрос – ответ – реакция, которая, вообще говоря, является открытой (т. е. пополняемой, изменяемой) во времени. Другими словами, знания, накопленные в предметной области, могут быть представлены открытым множеством логико-семантических сетей, упорядоченных по предметным темам. Задача предметной области может быть сформулирована в форме вопроса. Выявление в вопросе таких смыслов, как тема, содержание и объем вопроса, позволяет найти релевантные ЛСС, в которых могут содержаться как ответы, так и объяснения (реакции). Под предметной областью будем понимать область научно-практической деятельности человека, характеризуемую объектом и предметом исследования. Предметом исследования являются проблемы и задачи, связанные с объектом. Теории, методы, инструменты, опыт специалистов, научные и эмпирические знания и метазнания – ресурс, который используется специалистами для исследования проблем, продуцирования новых знаний, разработки теорий и технологий решения научных и практических задач. Будем считать, что накопленные знания предметной области представлены в научных отчётах, монографиях, статьях, учебных материалах, информационных фондах, справочниках, словарях и т. д. Будем считать, что информацию можно представить множеством тематических разделов, каждый из которых отражает определенный аспект знания предметной области. Каждой теме можно поставить в соответствие ЛСС ВОР. В этом случае знания предметной области формально можно представить следующим образом.

Обозначим *ПрОб* – наименование предметной области, $Тм_i$ – наименование i-й темы предметной области, *ЛССij* – наименование j-й ЛСС i-й темы. Тогда предметная область представляется как

$$ПрОб = \bigcup_{i=1}^{n} Тм_i;\ Тм_i = \bigcup_{j=1}^{m_j} ЛСС_{ij},$$

$$причём\ \bigcap Тм_i Тм_j \neq 0\ для\ \ i \neq j,$$

$$\bigcap ЛСС_{ij_i} ЛСС_{kr} \neq 0\ для\ \ i \neq k\ и\ j \neq r.$$

Представленные выше теоретические основы логико-семантических сетей ВОР позволяют описать предметную область множеством ЛСС, объединённых в тематические классы. ЛСС предметной области могут служить основой для создания поисковых информационных систем. Механизм поиска информации в предметной области на основе ЛСС может обеспечивать следующие режимы:

- движение по ЛСС, управляемое пользователем;
- поиск информации по заявке (вопросу).

4 Организация поиска в электронных библиотеках на основе ЛСС

Семантические поисковые системы на основе ЛСС ВОР могут иметь широкий спектр применимости, в том числе и в электронных библиотеках. Для реализации подобной системы в рамках конкретной ЭБ необходимо: построить множество ЛСС; реализовать механизм поиска информации в режиме ответа на вопрос; разработать навигационный механизм движения по ЛСС как вверх (от частного к общему), так и вниз (от общего к частному).

Построение множества ЛСС возможно на нескольких уровнях (слоях). Первый слой (самый нижний) содержит ЛСС конкретного информационного ресурса (документа). Следующим слоем может быть ЛСС, построенная на базе функционирующего в данной ЭБ тематического рубрикатора, и т. д. Таким образом, мы получим многоуровневый связанный набор графов, который обеспечивает поиск и навигацию в горизонтальном и вертикальном направлениях. Навигационный механизм здесь играет важную роль. Пользуясь им, пользователь получает возможность корректировать вопросы, на которые в системе не существует адекватных ответов. В режиме вопрос – ответ реализуется расчет меры близости вопроса, заданного пользователем, к уже существующим в ЛСС. Если такого вопроса нет, то он в дальнейшем может быть внесен в систему.

С точки зрения пользователя такая система позволяет в большинстве случаев найти ответ на поставленный вопрос. Пользователь задает вопрос и получает на него ответ с дополнительной информацией в виде реакций вопроса и ответа, которые помогают скорректировать вопрос либо воспользоваться уточняющими или обобщающими вопросами.

Создание, наполнение и сопровождение такой системы требует большой и серьезной работы, как технологической, так и организационной. Создание каталожной службы является трудоёмким ручным процессом. Поэтому для создания технологии формирования и поддержки каталога ЛСС требуется максимальная автоматизация, чтобы предоставить АРМ аналитикам, которые будут заниматься формированием ЛСС документов и предметных областей. При успешной реализации этой системы пользователям ЭБ будет предоставлена новая возможность – получать ответы на вопросы, заданные на естественном языке.

4.1 Имеющиеся наработки

Разработки ведутся в НИЦ Управления знаниями и распределёнными вычислениями Университета «Дубна». К настоящему времени:

- предложена методика и технология формирования ЛСС документа;
- разработано ПО для ввода, редактирования, накопления ЛСС в БД – прототип автоматизированного рабочего места (АРМ) разработчика ЛСС ВОР. Формы интерфейса АРМ разработчика ЛСС поисково-информационной консультативной системы (ПИКС) представлены на рис. 2; раздел «Тема» позволяет просматривать, находить и корректировать темы предметной области; раздел «ЛСС» дает возможность просматривать, редактировать и создавать ЛСС;
- разработаны методика поиска ответа на вопрос для корпуса документов и программное обеспечение (в частном случае) поисковой машины;
- на примерах осуществлена ручная апробация методик.

Методики апробированы в учебном процессе для различных дисциплин (Корпусная лингвистика, Теоретические основы автоматизированного управления, Системное моделирование и т. д.). В результате совместно с Технопарком г. Дубна в рамках проекта «Разработка портала «Содействие инновационной деятельности» разработана ЛСС «Поиск инвестора».

4.2 О возможности семантического поиска на основе ЛСС в архиве научных и научно-организационных документов ОИЯИ

В настоящее время многие научные и образовательные организации во всем мире создают собственные электронные репозитории (архивы)), размещая в них различные документы как научного, так и организационного характера и предоставляя к ним открытый доступ для всего мирового сообщества.

В зависимости от профиля организации эти архивы могут различаться тематической направленностью: фундаментальная или прикладная физика, астрономия, математика, химия, медицина и т. п. ОИЯИ является международным центром исследований в области физики частиц высоких энергий и физики атомного ядра. Однако, спектр тематических направлений этими двумя дисциплинами не исчерпывается и включает математику,

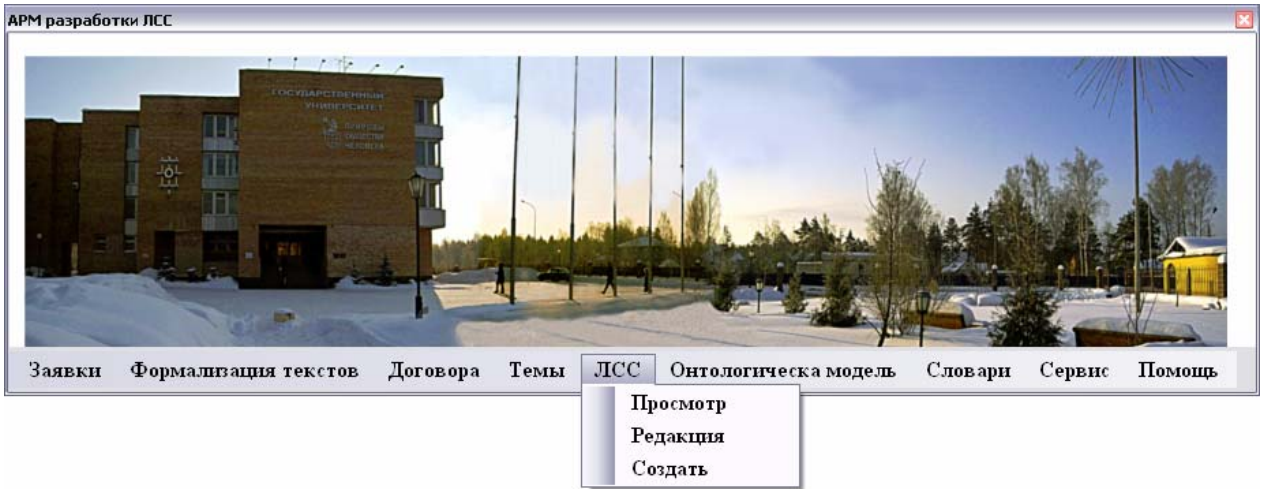


Рис. 2. Формы интерфейса АРМ разработчика ЛСС ПИКС

химию, прикладную физику, информационные и нанотехнологии. Созданный на базе библиотечного сервера JDS электронный архив содержит, кроме научных публикаций сотрудников ОИЯИ, являющихся основным типом документов, еще различные материалы, связанные с научно-организационной, педагогической и административной деятельностью. Типы документов, загружаемых в состав архива, помимо препринтов и статей в журналах включают диссертации, книги, годовые отчеты, материалы семинаров, тексты докладов, презентаций и материалы видеоконференций. Наличие развитого пользовательского интерфейса и необходимого библиотечного сервиса, обеспечиваемого пакетом CDS Invenio, превращает архив JDS в электронную библиотеку. Для повышения качества поиска нужной информации в архиве JDS представляется перспективным организация поиска на основе концепции ЛСС. Разнообразие тематических направлений и типов документов потребует формирования ЛСС для каждого направления, включающей описание предметных областей, возможность обработки и отображения химических и математических выражений, специализированных знаков и символов. Поскольку библиографические описания документов в JDS формируются с помощью метаданных, создание множества ЛСС в слоях выше первого (нижнего) облегчается возможностью машинной обработки данных.

5 Заключение

Существующие проблемы и разработки QA-систем пересекаются с проблемами вопросно-ответных систем для фондов-корпусов научной информации. Авторами предлагается создание «каталожной службы» и её поддержки для информационных фондов, создание вопросно-ответного навигатора, обладающего особенностями, обеспечивающие такие качества, как возможность в процессе поиска ответов уточнения вопросов, углубление понимания смысла вопроса, возможность в процессе поиска ответа уточнения, углубления и расширения знаний и как следствие получения нового знания.

Основной проблемой создания предлагаемой вопросно-ответной системы является максимальная автоматизация процесса создания и поддержки «каталожной службы» фонда, возможность осуществления такого проекта.

Методики анализа научной информации апробированы авторами в учебном процессе на различных дисциплинах.

Состоятельность и актуальность излагаемого подхода на данном этапе исследований и разработок подтверждается экспресс-анализом существующего положения QA-систем.

Литература

[1] Hamblin C.L. Questions// Australasian J. of Philosophy. – 1958. – V. 36. – P. 159-158.

[2] Аверьянов Л.Я. Почему люди задают вопросы? – М.: «Социолог», 1993.

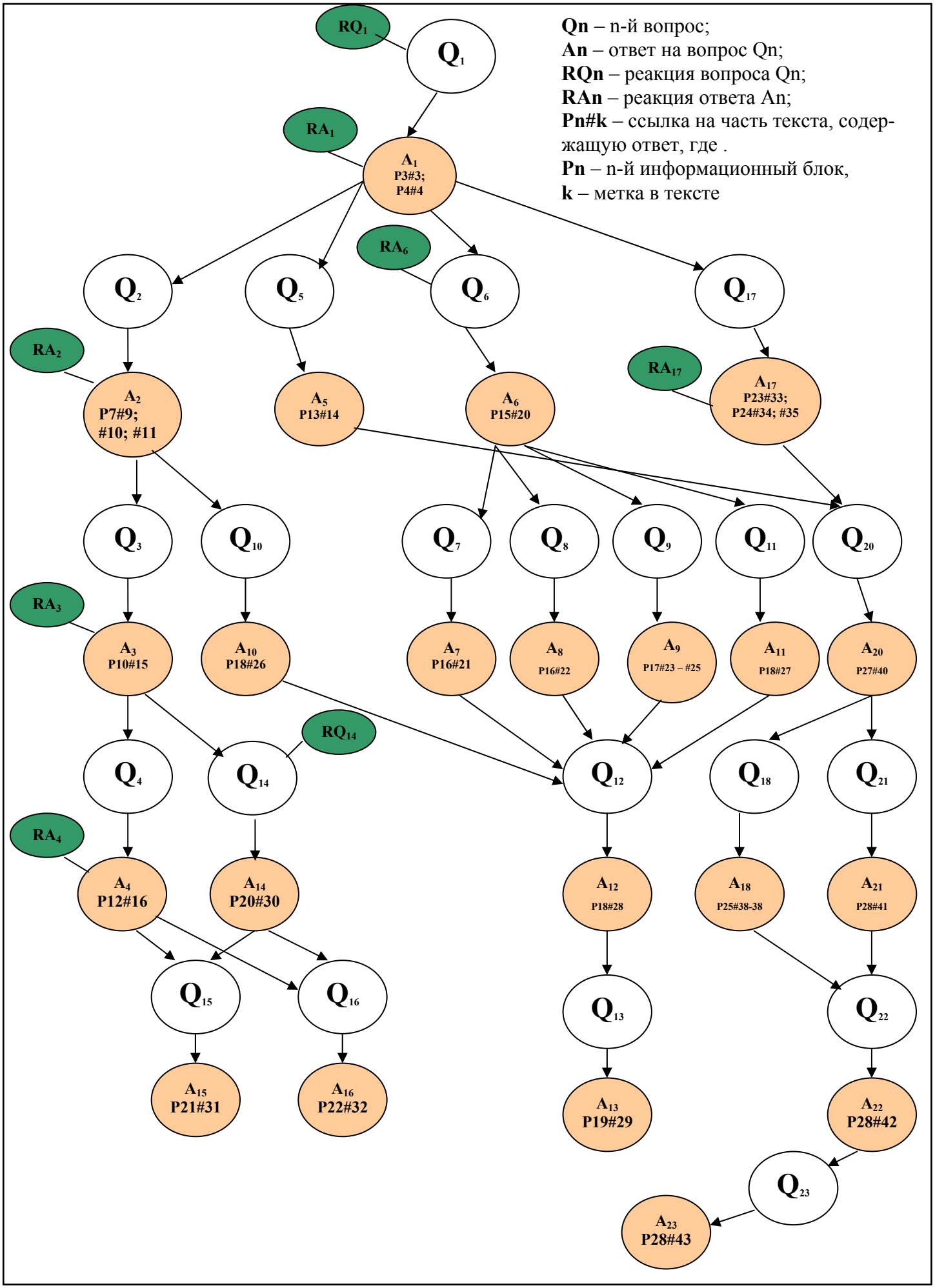
[3] Белнап Н., Стил Т. Логика вопросов и ответов. – М.: Прогресс, 1981. – 44 с.

[4] Добрынин В.Н., Лобачева М.В. Прототип семантической поисковой системы на основе логико-семантической сети «ВОПРОС – ОТВЕТ – РЕАКЦИЯ» // Электронный журнал «Системный анализ в науке и образовании», Международный университет природы, общества и человека «Дубна», кафедра САУ. – 2009, Вып.2. – http://www.sanse.ru/archive/11.

The search based on the logical semantic network "Question – answer –reaction"

V.N. Dobrynin, I.A. Filozova

The technology of semantic search in digital libraries based on the framework of Logical Semantic Network (LSN) "Question – response – reaction" is described. Such a system allows one to get an adequate response on the question, formulated in human language. The possibility of the usage of this approach for search and navigation in JINR digital archive is discussed.


Qn – n-й вопрос;
An – ответ на вопрос Qn;
RQn – реакция вопроса Qn;
RAn – реакция ответа An;
Pn#k – ссылка на часть текста, содержащую ответ, где .
Pn – n-й информационный блок,
k – метка в тексте
RQ1
Q1
RA1
A1 P3#3; P4#4
RA6
Q2
Q5
Q6
Q17
RA2
RA17
A2 P7#9; #10; #11
A5 P13#14
A6 P15#20
A17 P23#33; P24#34; #35
Q3
Q10
Q7
Q8
Q9
Q11
Q20
RA3
A3 P10#15
A10 P18#26
A7 P16#21
A8 P16#22
A9 P17#23 – #25
A11 P18#27
A20 P27#40
RQ14
Q4
Q14
Q12
Q18
Q21
RA4
A4 P12#16
A14 P20#30
A12 P18#28
A18 P25#38-38
A21 P28#41
Q15
Q16
Q13
Q22
A15 P21#31
A16 P22#32
A13 P19#29
A22 P28#42
Q23
A23 P28#43