КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2015 Т. 7 № 3 С. 661−673

Ки&М

СЕКЦИОННЫЕ ДОКЛАДЫ

УДК: 004.9

Технология формирования каталога информационного фонда

В. Н. Добрынин[1], И. А. Филозова[2, a]

[1] ГОУ ВПО «Международный университет природы, общества и человека «Дубна»,
Институт системного анализа и управления,
Россия, 141980, Московская обл., г. Дубна, ул. Университетская, д. 19
[2] Объединенный институт ядерных исследований,
Лаборатория информационных технологий,
Россия, 141980, Московская обл., г. Дубна, ул. Жолио-Кюри, д. 6

E-mail: [a] fia@jinr.ru

Получено 30 сентября 2014 г.

В статье рассматривается подход совершенствования технологий обработки информации на основе логико-семантической сети (ЛСС) «Вопрос–ответ–реакция», направленный на формирование и поддержку каталожной службы, обеспечивающей эффективный поиск ответов на вопросы [Большой энциклопедический словарь, 1998; Касавин, 2009]. В основу такой каталожной службы положены семантические связи, отражающие логику изложения авторской мысли в рамках данной публикации, темы, предметной области. Структурирование и поддержка этих связей позволят работать с полем смыслов, обеспечив новые возможности для исследования корпуса документов электронных библиотек (ЭБ) [Касавин, 2009]. Формирование каталога информационного фонда (ИФ) включает: формирование лексического словаря ИФ; построение дерева классификации ИФ по нескольким основаниям; классификация ИФ по вопросно-ответным темам; формирование поисковых запросов, адекватных дереву классификации вопросно-ответных тем (таблица соответствия «запрос → ответ ↔ {вопрос–ответ–реакция}»); автоматизированный поиск запросов по тематическим поисковым машинам; анализ ответов на запросы; поддержка каталога ЛСС на этапе эксплуатации (пополнение и уточнение каталога). Технология рассматривается для двух ситуаций: 1) ИФ уже сформирован; 2) ИФ отсутствует, его необходимо создать.

Ключевые слова: информационный фонд, Большие Данные, информационный поиск, пертинентность, навигация, информационно-поисковая система, семантические связи, логико-семантическая сеть «вопрос–ответ–реакция»

КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2015 Т. 7 № 3 С. 661−673

Ки&М

СЕКЦИОННЫЕ ДОКЛАДЫ

Cataloging technology of information fund

V. N. Dobrynin[1], **I. A. Filozova**[2]

[1] *Institute of System Analysis and Management, Dubna International University for Nature, Society, and Man*
19 Universitetskaya str., Dubna, Moscow region, 141980, Russia
[2]*Joint Institute for Nuclear Research, Laboratory of Information Technologies,*
6 Joliot Curie st., Dubna, 141980, Russia

The article discusses the approach to the improvement of information processing technology on the basis of logical-semantic network (LSN) Question–Answer–Reaction aimed at formation and support of the catalog service providing efficient search of answers to questions.

The basis of such a catalog service are semantic links, reflecting the logic of presentation of the author's thoughts within the framework this publication, theme, subject area. Structuring and support of these links will allow working with a field of meanings, providing new opportunities for the study the corps of digital libraries documents. Cataloging of the information fund includes: formation of lexical dictionary; formation of the classification tree for several bases; information fund classification for question–answer topics; formation of the search queries that are adequate classification trees the question–answer; automated search queries on thematic search engines; analysis of the responses to queries; LSN catalog support during the operational phase (updating and refinement of the catalog). The technology is considered for two situations: 1) information fund has already been formed; 2) information fund is missing, you must create it.

Keywords: information fund, Big Data, information search, pertinence, navigation, search engine, semantic relations, logic-semantic network «Question–Answer–Reaction»

Citation: *Computer Research and Modeling*, 2015, vol. 7, no. 3, pp. 661–673 (Russian).

Введение

Современные проблемы и задачи требуют для своего решения анализа больших объемов информации, распределенных в различных источниках. Время и ресурсы для решения проблем и задач, как правило, ограничены и зачастую несопоставимы с существующими механизмами поиска, селекции и аккумуляции требуемой информации по качеству. Неслучайно, что один из наиболее часто упоминаемых сегодня терминов в IT-области — это Big Data [Hilbert, López, 2011; Найдич, 2012; Якшонок, 2012].

Объемы научных фондов и их число растут с некоторой скоростью [Редькина, 2010]. Механизмы поиска становятся неэффективными по показателям «время», «деньги», «качество». С одной стороны, имеют место рост объема разнородной информации, рост в потребности в качественной информации. С другой — неэффективные поисковые информационные системы (ИПС) и общие вопросно-ответные системы. Специалисту в определенной области знаний важно иметь инструмент для эффективного исследования информации в массивах научных публикаций как основной продукции деятельности ученых и исследователей. В связи с этим возникает необходимость эффективных (высокий уровень релевантности, время поиска, большие объемы информации) ИПС и вопросно-ответных систем.

1. Данные–информация–знание

Существует множество взглядов на термин Big Data. McKinsey в отчете 2011 г. “Big data: The next frontier for innovation, competition, and productivity” определяет большие данные как такие объемы информации, которые выходят за рамки возможностей используемых в организации СУБД по их анализу и хранению. По мнению консорциума MIKE 2.0, термин означает не столько большой объем данных, сколько их сложность, вариативность, разнородность и неструктурированность. Источники Big Data разнообразны. Ими могут быть текстовые документы, файлы CAD-, CAM-приложений, показания датчиков, сенсоров, систем видеонаблюдения, системные журналы и пр. То есть это массивы данных, которые потенциально содержат ценную информацию, но в чем состоит ценность и как ее извлечь — непонятно. Таким образом, главными признакми данных категории Big Data можно назвать: а) затруднения в их обработке; б) трудность их интерпретации.

Характеристики 3V — объем/volume, скорость/velocity, многообразие/variety — требуют адекватных реакций информационных систем на решение задач поиска и обработки больших массивов с качеством Big Data. Это находит отражение на всех уровнях архитектуры современных систем: 1) аппаратное обеспечение; 2) системное и проблемно ориентированное обеспечение; 3) прикладное ориентированное обеспечение. Для решения задач такого рода активно задействуются параллельные распределенные архитектуры, грид, облачные инфраструктуры. Возможно, что появится совершенно новое решение в другой парадигме. Усиление аппаратной составляющей особенно актуально для систем массового обслуживания.

В характеристику «больших данных» *velocity* (динамичность, изменчивость, скорость изменения) может быть заложено изменение структуры данных. Хорошо структурированная информация может быть достаточно точно представлена данными. Слабо структурированная информация может быть представлена данными с высокой степенью неопределенности, что является следствием их изменчивости. Семантическое структурирование контента информационных фондов имеет целью формирование его смыслового поля и направлено на снижение степени неопределенности.

В настоящий момент информационные потребности пользователей направлены на получение новой информации и новых знаний из уже имеющихся массивов данных. Вычислительная нагрузка на компьютеры гораздо меньше, чем собственно обработка данных. То есть данные обрабатываются с целью получения информации, которую человек способен преобразовать в знание. Таким образом, актуальна цепочка «Данные–информация–знание».

Термин *знание* имеет несколько значений и толкований. Знание противопоставляется незнанию, т. е. отсутствию проверенной информации о чем-либо.

Знание — форма существования и систематизации результатов познавательной деятельности человека [Большой энциклопедический словарь, 1998]. Знание помогает людям рационально организовывать свою деятельность и решать проблемы, возникающие в ее процессе.

Знание (предмета) — уверенное понимание предмета, умение обращаться с ним, разбираться в нем, а также использовать для достижения намеченных целей.

Новое знание — совокупность сведений о существовании каких-либо объектов или их свойств, о процессах и явлениях действительности, ранее не известных науке и не входящих в существующую на данный момент систему человеческих представлений о мире [Касавин, 2009].

Пропущенное знание — знание, известное человечеству, но на данный момент не известное конкретному человеку (например, студенту, изучающему новый предмет образовательной программы).

Знание в широком смысле — субъективный образ реальности в форме понятий и представлений.

Знание в узком смысле — обладание проверенной информацией (ответами на вопросы), позволяющей решать поставленную задачу.

Знание в теории искусственного интеллекта (ИИ) и экспертных систем — совокупность информации и правил вывода (у индивидуума, общества или системы ИИ) о мире, свойствах объектов, закономерностях процессов и явлений, а также правилах использования их для принятия решений. Главное отличие знаний от данных состоит в их структурности и активности, появление в базе новых фактов или установление новых связей может стать источником изменений в принятии решений.

Целью научной деятельности является генерация нового знания, которому неизбежно предшествует информационный поиск (для изучения текущего состояния предметной области); обработка результатов поиска и генерация на ее основе новых данных. То есть прочтение, изучение любого научного текста происходит с определенной целью, формируемой в контексте задач, которые пользователи решают в процессе своей профессиональной деятельности. Знание фиксируется в определенных образах и знаках естественных и искусственных языков. Формой представления знаний может быть публикация различных жанров (статья, монография, препринт и т. д.), электронные архивы, репозитории, таблицы, базы данных и т. д.

2. Проблемы информационного поиска

Поиск — это сложная итерационная процедура, предполагающая уточнение запроса. Специалист в некоторой предметной области, осуществляющий поиск, имеет определенное представление о том, что из полученных результатов может являться ответом на его вопрос. Однако большинство классических информационно-поисковых и вопросно-ответных систем являются одноконтурными или двухконтурными (позволяющими осуществить поиск в найденном). Специалисту же удобнее уточнять запрос/вопрос не выходя из системы. Основные критерии оценки эффективности поисковых систем — скорость, точность и полнота ответов. Точность определяется тем, какая часть информации, выданной в ответ на запрос, является релевантной, т. е. относящейся к этому запросу. Полнота характеризуется соотношением между всей релевантной информацией, имеющейся в базе, и той ее частью, которая включена в ответ. Кроме этого, при оценке поисковых систем учитывается, с какими типами данных может работать та или иная система, в какой форме представляются результаты поиска и какой уровень подготовки пользователей необходим для работы в этой системе. Но для пользователя наиболее важна прагматическая характеристика информационного поиска, отражающая насколько результаты поиска удовлетворяют информационной потребности пользователя (пертинентность) — соответствие полученного результата информационной потребности пользователя независимо от того, как

полно и как точно эта информационная потребность была выражена с помощью информационного запроса или вопроса. Пертинентность измеряется степенью соответствия между ожиданиями пользователя и результатами поиска и определяется как отношение объема полезной для пользователя информации к общему объему полученной информации, найденной поисковой системой. Возможные причины низкой пертинентности информационного поиска изображены на рис. 1.

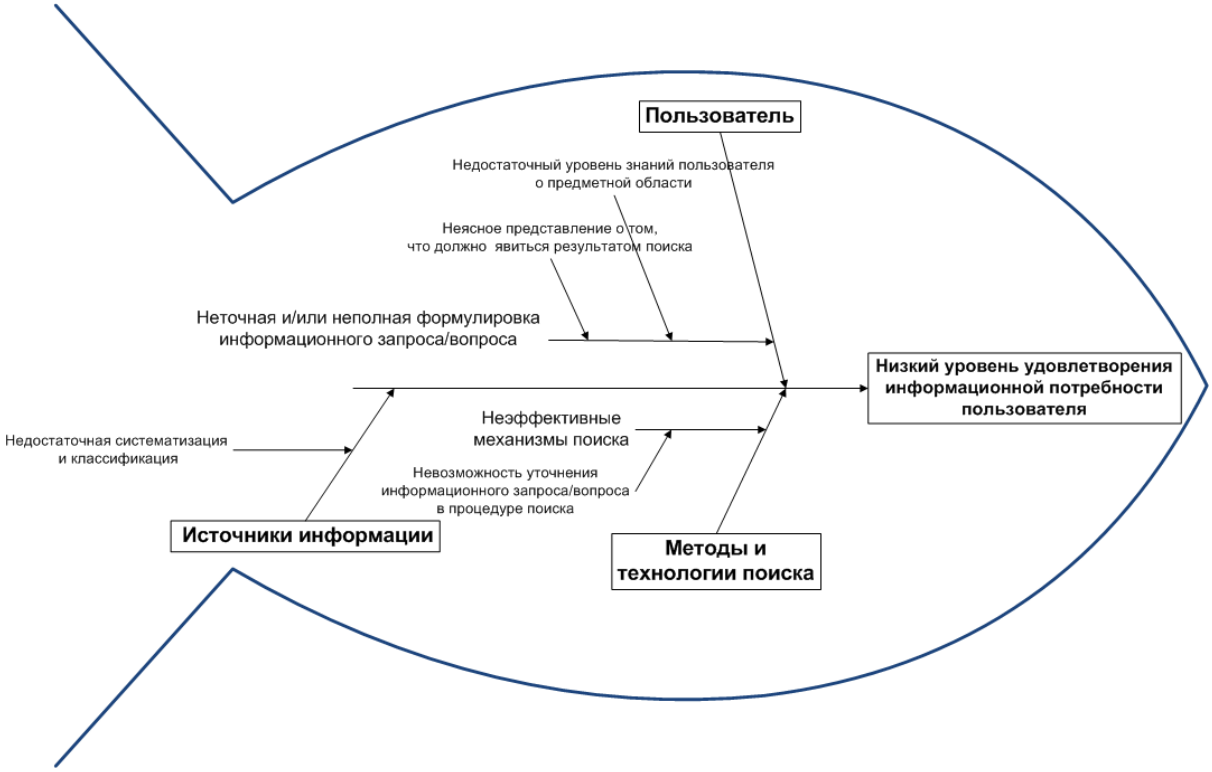

Рис. 1. Причины низкой пертинентности информационного поиска

Развитие методов организации информационного поиска подтверждает научно-практический интерес к решению этой проблемы.

3. Пути решения проблемной ситуации

Традиционные подходы к организации поиска информации можно разделить на три группы: методы индексного (или двоичного) поиска, статистические методы и методы, основанные на базах знаний. Для максимального удовлетворения информационных потребностей пользователей в настоящее время в поисковых системах широко применяются теории и методы семантических сетей, контент-анализа и интеллектуальный анализ текстов (Text Mining).

Индексный, или двоичный, поиск применяется главным образом со структурированными базами данных. Системы двоичного поиска имеют ограничения по точности, влияющие на возможность нахождения всей относящейся к запросу информации. В методах двоичного поиска не учитываются различные формы и значения слов; пользователю непросто угадать точные слова и фразы, которые были использованы авторами в документах. Системы двоичного поиска не могут также ранжировать документы по степени соответствия запросу, поэтому пользователь вынужден читать каждый документ, чтобы определить, насколько он соответствует запросу.

Статистические методы основываются на расчете различных частотных характеристик: частоты вхождения слова в документ, взвешенной частоты вхождения и частоты совместного

вхождения нескольких слов. Основной единицей информации, которой оперируют статистические методы, является отдельное слово, однако связи между словами рассматриваются исключительно с математической, а не с лингвистической точки зрения. В отличие от методов двоичного поиска статистические методы не требуют применения жесткого формального языка запросов. Они позволяют проводить ранжирование документов по степени соответствия запросу, что существенно повышает эффективность работы с поисковыми системами. Однако такие методы не всегда позволяют получить желаемые точность и полноту ответов, поскольку важность того или иного термина не напрямую связана с частотой его использования в документе.

Системы, основанные на базе знаний, занимаются поиском информации на основе некоторых внешних знаний. Они используют концептуальные отношения, которые не применяются при статистическом поиске. Одним из наиболее простых и распространенных способов представления знаний является файл синонимов. Другой подход к системам, основанным на базе знаний, использует иерархию терминов и понятий, создаваемую самими пользователями. Третий известен как подход на основе лингвистических правил.

Подход, использующий ссылочные документы, в том числе обычные словари и словари терминов, основан на смысловых значениях слов и называется семантической сетью. Как и словарь, семантическая сеть содержит множество определений для каждого хранимого слова. Однако определения родственных слов и понятий связываются между собой. Значения слов, наиболее подходящие для данного поиска, могут быть выбраны самим пользователем с целью повышения точности этого поиска. Подход на основе семантических сетей реально объединяет статистический поиск и поиск на основе базы знаний. При этом используются смысловые значения слов для определения и классификации отношений, которые статистический поиск не отслеживает.

Системы, основанные на базах знаний, гораздо удобнее тех, которые базируются на двоичном поиске. Однако сегодня лишь подход, основанный на построении семантических сетей, свободен от ограничений, присущих двоичному поиску; он обладает достаточной гибкостью, доступен для расширения и не слишком громоздок при эксплуатации.

Под контент-анализом в интернет-поисковиках понимают оценку структуры и материалов веб-ресурса с точки зрения поисковой оптимизации. При контент-анализе оцениваются такие смысловые единицы, как наличие контента (страниц), релевантного поисковому запросу; уникальность контента; удобство использования (дизайн, структура сайта, навигация); удобство контента для восприятия; качество html-кода.

Text Mining представляет собой множество методов обработки текста, в результате применения которых появляются новые, ранее не предполагавшиеся знания. Это междисциплинарная область, в которой используются базовые технологии Data Mining совместно с методами информационного поиска, извлечения информации, математической лингвистики, создания онтологий, классификации, кластеризации и др.

4. Логико-семантические сети «вопрос–ответ–реакция»

Работа специалиста-профессионала с информационными фондами предполагает наличие системы каталогизации и классификации материала. В зависимости от специализации контента информационные системы обеспечиваются электронными каталогами с целью описания ресурсов для их однозначной идентификации и обеспечения доступа к ним. В рамках заданной проблемной темы предлагается технология формирования и поддержки *каталожной* службы, которая обеспечивает эффективный поиск ответов на вопросы. Основой такой *каталожной* службы является упорядоченное открытое множество логико-семантических сетей (ЛСС) «вопрос–ответ–реакция» [Добрынин и др., 2014].

Любая научно-практическая область знаний включает предмет исследования, который может быть представлен проблемным полем (перечнем проблемных вопросов), являющимся основой для научной и практической деятельности. Проблемные вопросы могут быть пред-

ставлены 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 определенная реакция. Ответы могут порождать, в свою очередь, вопросы. Таким образом, проблемный вопрос соотносится с определенной темой 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 (ЛСС), упорядоченных по предметным темам. Задача предметной области может быть сформулирована в форме вопроса. Выявление в вопросе таких смыслов, как тема вопроса, содержание вопроса, объем вопроса, позволяет найти релевантные ЛСС, в которых могут содержаться как ответы, так и необходимые объяснения (реакции). Ввод реакций помогает пользователю понять, получил ли он релевантный и пертинентный ответ на свой вопрос. В качестве реакций может выступать дополнительная информация по теме вопроса и ответа, иллюстрации, изображения, таблицы, ссылки на сайты, словари, рубрикаторы, каталоги и т. д. Такими реакциями может сопровождаться как вопрос, так и ответ, что позволит пользователю лучше и быстрее сориентироваться в предметной области.

4.1. Основные положения ЛСС

Под логико-семантической сетью будем понимать множество вопросов, ответов и связей между ними, образующее целостную систему. Под целостностью ЛСС имеется в виду следующее:

1) множество «вопрос–ответ» относится к определенной теме предметной области;

2) множество «вопрос–ответ» иерархически упорядочено по принципу «от общего к частному»;

3) на нечетном уровне иерархии находятся вопросы, на четном уровне — ответы;

4) вопросы *i*-го уровня иерархии связаны только и только с ответами *i* + 1-го уровня;

5) вопросы *i*-го уровня связаны с ответами *i* – 1-го уровня;

6) вопрос *i*-го уровня семантически связан с ответами i+1-го уровня если удовлетворяет условиям А или В. В случае удовлетворения условию А имеет место конечная вершина; В случае удовлетворения условию В из данного ответа следуют вопросы *i* + 2-го уровня;

7) на *i* = 1 уровне находятся вопросы, которые раскрываются множеством ответов *i* = 2-го уровня, частично или полностью охватывающее тему предметной области;

8) на *i* = 3-м уровне находятся вопросы, которые восполняют и уточняют ответы *i* = 2-го уровня.

Единицей ЛСС является логическая связка «вопрос–ответ» и связанные с ними реакции. Вопросы всегда опираются на уже известное знание, выступающее их базисом и выполняющее роль предпосылки вопроса. Постановка вопроса и поиск информации для формирования ответа составляют вопросно-ответную логическую форму развития знаний. Таким образом, ЛСС «вопрос–ответ–реакция» можно представить в виде направленного графа (рис. 2). Суть излагаемого подхода состоит в том, что любая задача или научно-технический текст может быть представлен в виде логической последовательности вопросов и ответов, которая дополняется полезной информацией.

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, характеризующее предпосылки вопроса и область поиска ответа. Реакция позволяет учитывать и использовать дополнительные знания о предметной области.

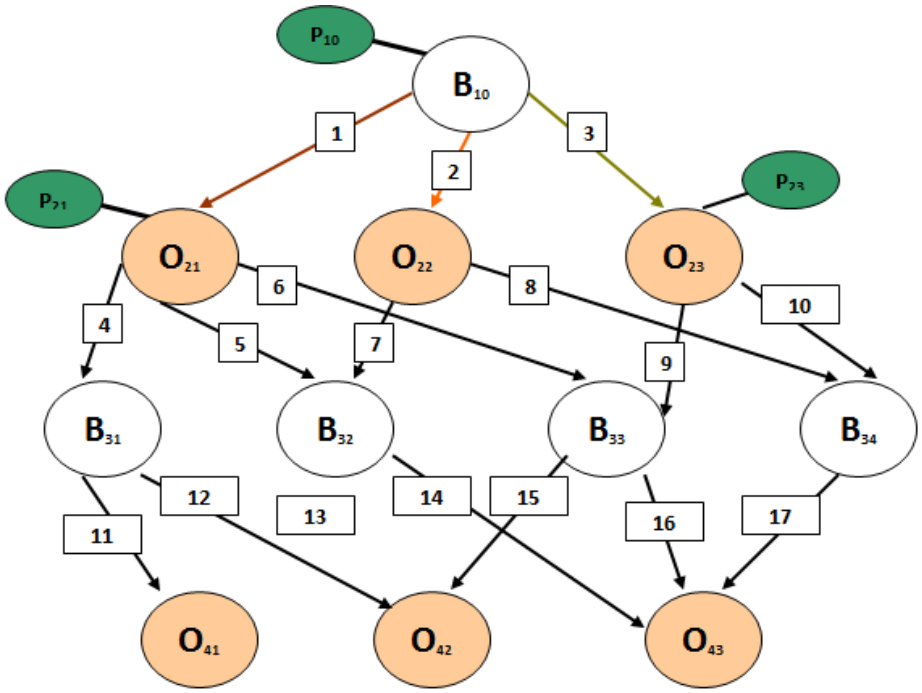

Рис. 2. Граф «вопрос–ответ–реакция»

Типы реакций

1. Реакции вопроса — это описание области предпосылки вопроса (для осознания обстоятельств и причин возникновения вопроса и дальнейшего установления смыслового соответствия с областью ответа). Реакция вопроса характеризует область вопроса — смысловое пространство, из которого аргументируется возникновение вопроса (хотя самой аргументации нет).

2. Реакции ответа — это описание области ответа (для осознания смысла вопроса и смысловой связи с ответом). Реакция ответа — смысловое пространство, имеющее связь с пространством вопроса, из которого следует ответ. Каркасом связи вопроса и ответа является ЛСС (рис. 3–5).

5. Электронные научные фонды & BigData

Сегодня в мире издается примерно 25 000 научных журналов, публикующих 1 млн статей в год, что соответствует ежедневному выпуску порядка 2700 публикаций [Редькина, 2010]. Для изучения таких объемов информации стандартными поисковыми технологиями специалисту потребуется значительное количество времени. Соответственно, важно иметь инструмент для эффективного исследования информации в массивах научных публикаций как основной продукции деятельности ученых и исследователей.

В электронных научных фондах количественные характеристики меняются быстро, качественные характеристики меняются медленнее, чем у сетевых систем. Качественное изменение научно-практических, технологических, технических знаний происходит скачкообразно и взаимно обусловлено и до определенной степени непредсказуемо. Можно считать, что на определенных временных периодах эти массивы знаний неизменны. В такие периоды исследователи работают традиционно, придерживаясь существующей парадигмы. А на периодах качественного перехода изменяются парадигмы и возникают новые знания, суть которых состоит в проявлении определенных косвенных качественных тупиковых (предельных) возможностей используемых знаний без желаемых эффектов. Такого рода деятельностью занимаются специалисты — эксперты, определяющие возможности будущих направлений научных, технологических и технических исследований и достижений.

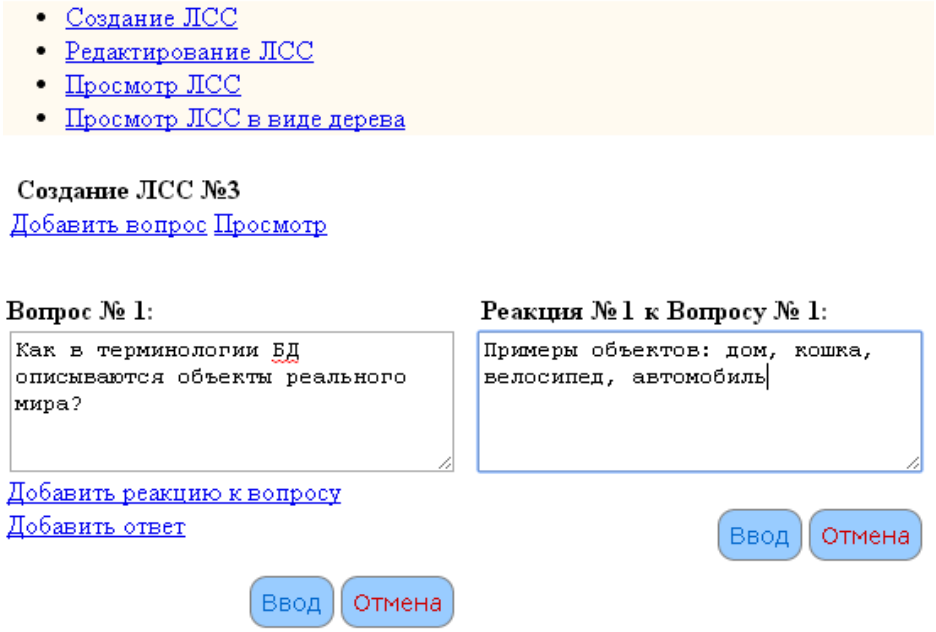

Рис. 3. АРМ аналитика. Режим создания ЛСС: формирование вопроса и реакции

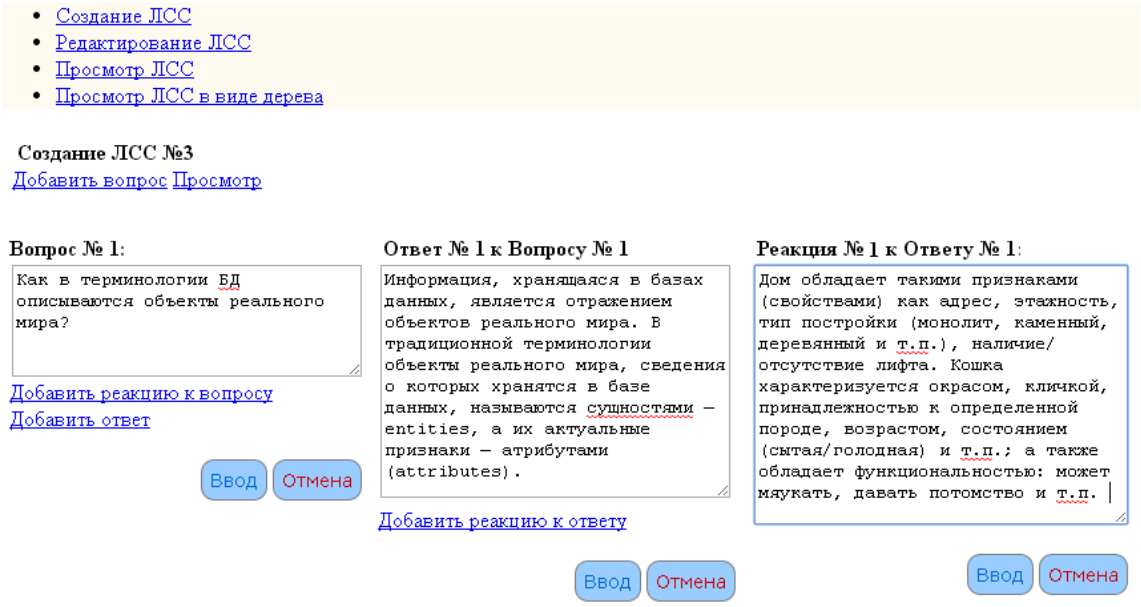

Рис. 4. АРМ аналитика. Режим создания ЛСС: формирование ответа и реакции

Но грань между пропущенными и новыми знаниями не очевидна. Ученые и исследователи часто прибегают к рассуждениям по аналогии. Аналогия — мощный инструмент в науке, обеспечивающий генерацию новых идей, гипотез и решений. По сути, аналогия является переносом, т. е. понятия, допущения, модели переносятся из одной области человеческого знания, где они показали свое эффективное применение, в другую область, в которой исследователь пытается разрешить некую проблемную ситуацию. Перенос идеи, уже успешно апробированной в другой области, подкрепляет убежденность ученых в эффективности используемых методов. С помощью аналоговых переносов устанавливаются взаимосвязи между новыми идеями и тем, что уже считается достоверным знанием. Есть мнение, что новое знание — это знание, которое

не имеет аналогии. Но история науки иллюстрирует, что самое радикальное новшество, как правило, проявляет неожиданные аналогии с уже имеющимися знаниями [Ивин, 2002].

Рис. 5. АРМ аналитика. Режим редактирования ЛСС: изменение текста реакции

Так, современная биология активно использует математический аппарат, в частности теорию дифференциальных уравнений, теорию вероятностей и статистику, теорию игр для формализации представлений о структуре, принципах функционирования и взаимоотношений живых организмов. Например, в популяционной динамике возникла математическая теория взаимодействия популяций одного трофического уровня (конкуренция) или разных трофических уровней (хищник–жертва), в которой сложные живые системы описываются при помощи систем обыкновенных дифференциальных уравнений. В математике дифференциальные уравнения не являются новым знанием, но перенос их в биологию позволяет не только моделировать, но и прогнозировать процессы в живых системах, что может привести биологов к новым знаниям — понимаю закономерностей соответствующих биологических процессов.

Логично предположить, что специалист, решающий некоторую профессиональную задачу в определенной области человеческого знания, будет действовать по принципу аналогового переноса, а именно, попытается выяснить, какие существуют успешные решения подобных задач. То есть первый этап его деятельности — информационный поиск. Из этого следует, что в системе поиска информации надо уделить внимание поиску пропущенных и/или новых знаний.

6. Направления исследования

Роль вопроса в процессе познания чрезвычайно важна. Совокупность вопроса и ответа формирует *единицу мысли*. В форме вопроса осуществляется постановка новых проблем в науке, с помощью вопросов люди получают новую информацию в социальной практике. Соответственно, любая задача может быть сформулирована в виде вопроса, а ее решение представлено как серия взаимосвязанных вопросов и ответов.

Разработка метода и механизма эффективного поиска множества релевантных ответов на вопрос включает:

1) разработку технологии формирования и поддержки *каталожной* службы информационного фонда, обеспечивающей эффективный поиск ответов на вопросы, на основе ЛСС «вопрос–ответ–реакция»;

2) создание инструментария (ПО) — АРМ аналитика для структурирования информационного фонда, предназначенного для создания и редактирования множества ЛСС.

Основой метода является способ описания научно-технической и образовательной информации множеством логико-семантических сетей «вопрос–ответ–реакция». Основой механизма поиска является способ движения по ЛСС, управляемый пользователем посредством выбора в ЛСС узлов — вопросов или ответов — на основе онтологической модели вопроса пользователя.

В рамках заданной проблемной темы предлагается технология формирования и поддержки *каталожной* службы, которая обеспечивает эффективный поиск ответов на вопросы. Стержнем такой *каталожной* службы является упорядоченное открытое множество логико-семантических сетей (ЛСС) «вопрос–ответ–реакция» [Добрынин, Филозова, 2010; Добрынин, Филозова, 2014]. С помощью специализированного навигатора специалист-профессионал, выполняющий поиск, может либо уточнять вопрос, либо его углублять, получая соответствующие связки «вопрос–ответ». Эта возможность достигается за счет введения Реакции, позволяющей учитывать и использовать дополнительные знания о предметной области. Тем самым пользователь от имеющихся знаний может получить расширенные знания, углубленные знания, уточненные знания или пропущенные знания. При этом за счет 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.

Данный подход позволяет заменить неопределенности, связанные с информационным поиском (когда не ясно точно, какая информация ищется и с какой целью) на более продуктивную технологию, ориентированную на пространство смыслов. Поисково-обрабатывающая система на основе ЛСС — это еще один путь совершенствования информационных технологий обработки информации.

7. Разработка АРМ аналитика для структурирования информационного фонда

Создание, наполнение и сопровождение такой информационной системы требует большой и серьезной работы, как технологической, так и организационной. Некоторую ее часть можно автоматизировать, предоставив соответствующее программное обеспечение аналитикам — АРМ аналитика для создания и редактирования множества ЛСС.

Формирование ЛСС базируется на методике анализа научных текстов, согласно которой текст исследуется каталогизатором с точки зрения [Filozova, 2012]:

1) смыслового соответствия заглавия и содержания;

2) набора фильтров:

F1 — общая часть: анализ проблемы, ее история, обзор, актуальность;

F2 — авторские понятия: вводимые авторами новые термины, обще употребляемые термины с авторской интерпретацией, сужающие семантику;

F3 — примеры и иллюстрации;

F4 — идея автора: описание и раскрытие основной авторской идеи;

3) формирования базовых вопросов, на которые отвечает текст.

На полученном таким образом материале далее строится ЛСС информационного ресурса: формулируются вопросы, ответы и реакции к ним.

В соответствии с вышесказанным АРМ аналитика предназначен для создания и редактирования множества ЛСС и обеспечивает следующую функциональность: 1) создание ЛСС (рис. 3, рис. 4); 2) редактирование ЛСС (рис. 5); 3) просмотр ЛСС (рис. 6).

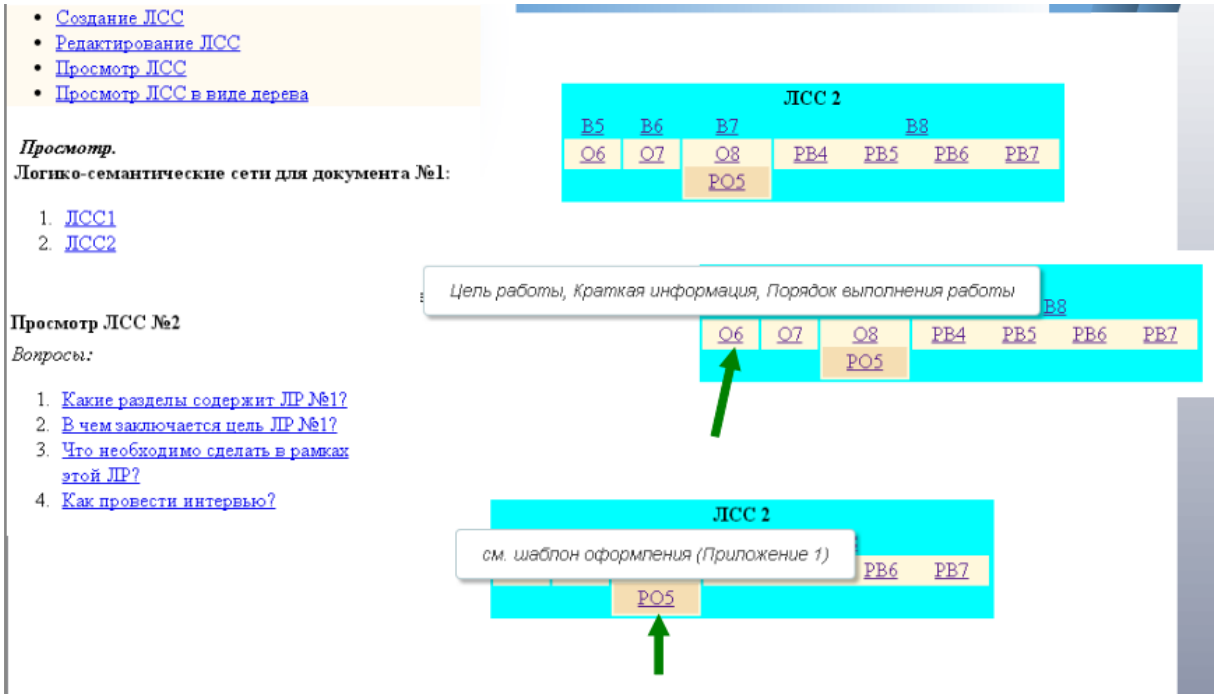

Рис. 6. АРМ аналитика. Режим просмотра ЛСС

Заключение

Одной из важных прикладных проблем эффективного поиска информации в условиях жестких временных ограничений — это проблема поиска новых и/или пропущенных, уточненных, углубленных знаний в точках бифуркации социотехнических систем. Информационные фонды (в том числе и электронные библиотеки), отражающие накопленные знания (теоретические, прикладные, прагматические), являются источниками генерации новых идей и формирования постановки и решения широкого спектра задач: исследование, экспертиза, инженерная задача, конструкторская задача и пр. Поиск необходимых для решения поставленной задачи знаний (в узком смысле) в некотором массиве информации может быть осуществлен посредством вопроса на языке предметной области.

В рамках заданной проблемной темы предлагается технология формирования и поддержки *каталожной* службы, которая обеспечивает эффективный поиск ответов на вопросы. Стержнем такой службы является упорядоченное открытое множество логико-семантических сетей «вопрос–ответ–реакция». Механизм навигации позволяет уточнять вопрос либо его углублять, получая соответствующие связки *вопрос–ответ*. Эта возможность достигается за счет введения *реакции,* позволяющей учитывать и использовать дополнительные знания о предметной области. Тем самым пользователь от имеющихся знаний может получить расширенные знания, углубленные знания, уточненные знания или пропущенные знания. При этом за счет *реакции* пользователь может контролировать согласованность смыслового собственного понимания вопросов и ответов и понимания вопросов и ответов, заложенных в семантической поисковой системе. Поскольку система открытая (пополняемая, изменяемая во времени), пользователь в процессе взаимодействия с системой может уточнять и расширять саму ЛСС. То есть пользователь при активном развитии системы становится соавтором смыслового пространства ЛСС. В этом состоит адаптация системы. Таким образом, поисково-обрабатывающая система на ос-

нове ЛСС — это еще один путь совершенствования информационных технологий обработки информации.

Список литературы

Большой энциклопедический словарь. 2-е изд., перераб. и доп. — М.–СПб.: Большая российская энциклопедия, 1998. — 1456 с.

Добрынин В. Н., Филозова И. А. Поиск в научной электронной библиотеке на основе логико-семантической сети «вопрос–ответ–реакция» // Труды XII Всероссийской научной конференции RCDL'2010 «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». — Казань: Казанский университет, 2010. — С. 301–308. — Библиогр.: С. 308. — ISBN: 978-5-98180-838-8

Добрынин В. Н., Филозова И. А. Семантический поиск в научных электронных библиотеках // Информатизация образования и науки. — 2014. — № 2(22). — С. 110–110.

Ивин. А. А. Логика. Учебник для гуманитарных факультетов. — М.: ФАИР-ПРЕСС, 2002.

Касавин И. Т. Энциклопедия эпистемологии и философии науки. — М.: «Канон+», РООИ «Реабилитация», 2009.

Найдич А. Big Data: проблема, технология, рынок // КомпьютерПресс №1. 2012 [Электронный ресурс]. URL: http://www.compress.ru/article.aspx?id=22725&iid=1044

Редькина Н. С. Современное состояние и тенденции развития информационных ресурсов и технологий // Библиосфера. — 2010. — № 2. — С. 23–29.

Якшонок Г. Эффективный поиск и анализ научно-исследовательской информации в SciVerse: Scopus, Hub, ScienceDirect // МГИМО, 2012. [Электронный ресурс]. URL: http://mgimo.ru/files2/y03_2012/220642/MGIMO_March-2012.ppt

Filozova I. A. Technology of semantic structuring of the digital library content // Proceedings of the 5th International Conference "Distributed Computing and Grid-technologies in Science and Education". Dubna: JINR, 2012. — P. 117–122.

Hilbert M., López P. The World’s Technological Capacity to Store, Communicate, and Compute Information // Science. — April. — 2011. — Vol. 332, no. 6025. — P. 60–65. — DOI: 10.1126/science.1200970.