УЧЕНЫЕ ЗАПИСКИ КАЗАНСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА
Том 151, кн. 3 Физико-математические науки *2009*

УДК 002.6:025.3/.4

СПЕЦИАЛИЗИРОВАННАЯ ФАКТОГРАФИЧЕСКАЯ XML-ОРИЕНТИРОВАННАЯ ИНФОРМАЦИОННАЯ СИСТЕМА ДЛЯ ХРАНЕНИЯ И ПОИСКА ЛИТЕРАТУРНЫХ ДАННЫХ

С.М. Хусаинов, Л.Ю. Ильина, Антон О. Кузьмин, Андрей О. Кузьмин, В.Н. Пармон

Аннотация

В статье предложен подход к решению проблемы организации работы научного коллектива с хранилищем своих литературных и фактографических данных, а также к организации поиска в собственных текстовых коллекциях или в сети Интернет с использованием тезаурусов и рубрикаторов и поискового сервиса Яндекс.

Ключевые слова: обработка литературных данных, поиск в текстах, тезаурус по катализу, рубрикатор по катализу, наука о катализе.

Введение

В современном мире стремительное увеличение объемов производимой информации, накопление данных и знаний во всех областях естественных наук, а также необходимость быстрой разработки новых технологических решений на базе существующих знаний заставляют искать новые пути организации и интенсификации работы научных коллективов, экспертов, руководителей и менеджеров научных и технических проектов.

В частности, такие сферы научно-технической деятельности, как катализ, химия, химическая технология и т. д., представляют собой сложный сплав из самых различных дисциплин, скорость производства информации в которых достигает огромных масштабов.

Информационные системы, облегчающие профессиональную деятельность научного (и/или технического) коллектива, начали создаваться еще в середине прошлого века, выполняя различные (актуальные для того времени) задачи. Так, можно привести примеры следующих широко распространенных типов систем.

Фактографические базы данных – обычно носят информативный характер, содержат общие данные об объектах и их свойствах, выраженные в краткой (числовой) форме.

Библиографические базы данных – содержат весьма крупные объекты: обзоры, книги, публикации, относящиеся к какой-то одной области деятельности.

Электронные библиотеки – распределенные информационные системы, позволяющие надежно сохранять и эффективно использовать разнородные коллекции электронных документов через глобальные сети.

Порталы доступа к полнотекстовым источникам – WEB-сайты, организованные как многоуровневое объединение различных ресурсов и сервисов, предоставляющие доступ к библиографическим базам данных и электронным библиотекам с помощью WEB-интерфейса.

Системы обработки и анализа данных – позволяют проводить определение критериев отбора достоверных данных, выборку данных, создание набора хранимых объектов. В некоторых случаях для этого механизма достаточно вполне очевидного механического отбора данных, в других же необходимо применение экспертных систем.

LIMS (Laboratory Information Management System) – «Лабораторные информационно-управляющие системы». Это могут быть исследовательские лаборатории, отделы контроля качества предприятий, лаборатории обеспечения качества и др. Возможности LIMS определяются областью их применения – каждая конкретная LIMS предназначена для решения определенного узкого класса задач.

В действительности, несмотря на наличие хорошо изученных и формализованных способов организации информации, из-за возрастающего объема становится всё более сложным организовывать эффективный поиск, систематизацию, обработку данных, содержащихся в традиционных источниках данных – книгах, статьях, патентах, дневниках и т. д. В то же время за относительно небольшое время существования информационно-коммуникационных технологий накоплен очень большой объем разнообразных данных, представленных исключительно в электронной форме.

Анализ текстовой информации требуется для получения базовой информации о текущем уровне данного научно-технического направления, поиска конкретных подходов к решению поставленных задач, для определения приоритетных направлений, перспективных тематик, технологий и материалов, которые будут востребованы в ближайшем будущем. В частности, проведение полноценного анализа всех доступных полнотекстовых источников имеет особое значение при проведении прикладных исследований для успешной реализации целевого процесса.

Однако извлечение и анализ информации из таких источников традиционными «ручными» методами требует больших временных затрат высококвалифицированных экспертов. Очень высокие трудовые и временные затраты идут на подборку и создание коллекций текстовых документов, относящихся к определенной проблематике, а также на последующую выборку и анализ необходимых текстовых данных. Кроме того, имеющаяся в литературе научно-техническая информация характеризуется наличием сильного «информационного шума», затрудняющего выбор значимой, достоверной и, следовательно, полезной информации.

Таким образом, частью общей проблемы является необходимость наличия возможности проведения «интеллектуального» поиска данных в текстовых источниках при существенно сниженных требованиях к пользователю, касающихся его умения работать со сложными поисковыми запросами. При этом поисковая система (настраиваемая на работу в конкретной предметной области) должна путем корректировки и дополнения запроса пользователя обеспечивать получение максимального числа релевантных результатов поиска.

Действительно, в настоящее время поиск осуществляется в основном по ключевым словам без учета содержания в текстах ассоциативных и семантических связей с той или иной областью знания. Суть проблемы состоит в том, что обычные методы поиска по ключевым словам зачастую не дают нужный результат. Это заключается в большом количестве «мусора» и игнорировании некоторых действительно нужных пользователю результатов. Одна из причин заключается в том, что на простые текстовые запросы данные системы выдают документы, которые содержат эти ключевые слова, но не обязательно соответствуют реальной информационной потребности пользователя, причем нужный результат может затеряться среди других, более релевантных, но ненужных пользователю результатов. Другая причина – задача построения грамотного запроса в данной предметной

области с помощью ключевых слов может быть слишком сложной для пользователя. В результате пользователь, введя запрос из 2–3 слов, получает либо слишком много вариантов для выбора и при этом ценная информация не может быть найдена в груде мусора, либо не находит нужную информацию по причине составления недостаточно проработанного запроса, ввиду незнания синонимов используемых терминов и т. д.

Основной и базовый метод поиска сегодня – поиск по образцу. Именно он используется в широкодоступных поисковых системах, таких, как Яндекс и Google, и повсеместно распространен. Его главный недостаток для всех очевиден: поисковик выдаёт зашумленный случайными совпадениями результат, а ссылки на документы не соответствуют контексту запроса. Соответственно, встает задача уменьшения шума, а также задача «угадывания» того, что же пользователь искал на самом деле.

В настоящее время разрабатываются теоретические подходы к решению указанных проблем, например, предлагается использовать связку тезаурусов и онтологий для интеллектуальных систем автоматической обработки текста на естественных языках [1–3]. Отражено несколько попыток построения предметных онтологий для организации порталов знаний. Предложены подходы к применению онтологий для организации поиска информации в больших массивах знаний и данных, в частности в сети Internet.

Таким образом, отсутствие специализированных компьютерных систем для хранения коллекций текстовых научных данных и поиска в них содержательной информации сильно затрудняет работу научных экспертов при принятии тех или иных решений в современных динамично развивающихся областях науки и техники. На сегодняшний день актуальность указанных проблем несомненна, но даже при наличии существующих мощных аппаратных и программных средств она так и не решена должным образом.

Целью настоящей работы, проводимой на базе Института катализа им. Г.К. Борескова СО РАН, является разработка специализированной информационной системы хранения и анализа коллекций текстовых научных данных научного коллектива, а также подсистемы интеллектуального поиска в созданных литературных коллекциях или в сети Интернет.

1. Информационная система хранения научных данных

Данная XML-ориентированная система предназначена для сопровождения научных исследований в области естественных наук, сохранения их результатов и последующего анализа накопленных данных. Работа с системой осуществляется через веб-интерфейс.

Цель создания данной системы можно пояснить на следующих примерах.

Каждый научный сотрудник, ведущий одну или несколько исследовательских тематик, обязательно сталкивается с проблемой структурирования имеющихся у него литературных данных, ссылок, высказываний, мнений и собственных заметок, относящихся к собранным литературным данным. Поиск нужной информации в собственных архивах и архивах своих подчинённых порой представляет собой весьма трудоёмкую задачу. Помимо этого существует проблема получения содержательной информации из чужих архивов научной литературы, в том числе своих сотрудников, которая связана с различиями в подходах к обработке, классификации и структурированию содержащейся в них информации, а также необходимость сохранения и передачи накопленных знаний.

Типичная ситуация – ушедший по тем или иным причинам из организации эксперт, после которого осталось несколько сотен оттисков, заметок и т. д. Сами по себе они не могут являться консультативной системой, поскольку человеку, чтобы разобраться в них, придется пройти тот же путь, что и предыдущему эксперту, пусть несколько более короткий, но также очень длинный. Это происходит оттого, что потеряны все знания, которые представлял собой эксперт, то есть ассоциативные связи, структура объектов, с которыми имеет дело данная область, и иерархия связей между ними. Действительно, известно, что лучше поговорить полчаса с экспертом, специалистом в данной области, чем просидеть в библиотеке целый месяц. В связи с естественной миграцией и уходом людей гигантские усилия организаций тратятся на воссоздание знаний, накопленных раннее.

Следующим характерным примером является лабораторный журнал экспериментатора, утрата которого для организации может являться фатальной. При этом зачастую даже сам экспериментатор с трудом разбирается в своих прошлогодних записях (не говоря о его коллегах), а координация работы целого экспериментального коллектива требует значительных усилий. Нередко люди выполняют одну и ту же работу в разное время или в разных местах из-за утраты или недоступности данных, полученных своими коллегами.

В рамках данной информационной системы (ИС) разработан эффективный способ для решения проблемы предоставления пользователям возможности для формализованного описания, хранения и структурирования научных данных, в первую очередь текстовых, основанный на применении XML-технологий. Реализуемый подход обеспечивает не только эффективное структурирование и хранение разнородных данных с возможностью организации их последующей обработки, но и предоставляет возможность для сохранения экспертных оценок и формализации знаний в некоторой предметной области.

Основная идея использования XML состоит в возможности произвольного моделирования структур хранимых данных с использованием XML-схемы. Кроме того, эксперт может устанавливать между схемами взаимные связи (обычным для языка XML-схем образом), что позволяет повторно использовать одни и те же структуры данных для моделирования новых.

На основе предлагаемого подхода к хранению данных, в базе данных определены дополнительные логические объекты, позволяющие эффективно обрабатывать литературные данные. Система позволяет сохранять экспертные знания в виде метаинформации о накопленных данных, создавать и сохранять литературные коллекции, посвященные конкретной научной проблеме.

Ниже представлен краткий список основных задач, поставленных перед системой, в рамках работы с коллекциями литературных данных (публикаций) (рис. 1):

- возможность определения пользователем структуры данных публикации (заголовок, содержание, глава и т. д.) и сохранения её метаинформации (автор, дата, ключевые слова и т. д.), а также возможность хранения неструктурированной публикации;
- выставление типизируемых пользователем оценок, заметок и примечаний к конкретной публикации (например, «не заслуживает доверия», «важно» и т. д.);
- классификация публикаций по темам (создание тематического рубрикатора);
- установление взаимных ссылок между публикациями для отслеживания структуры взаимного цитирования документов;
- хранение коллекций структурированных тематических ссылок на выделенные места в публикациях, содержащих важную информацию. (Данные ссылки по сути представляют собой «базу данных» выдержек из публикаций, например,

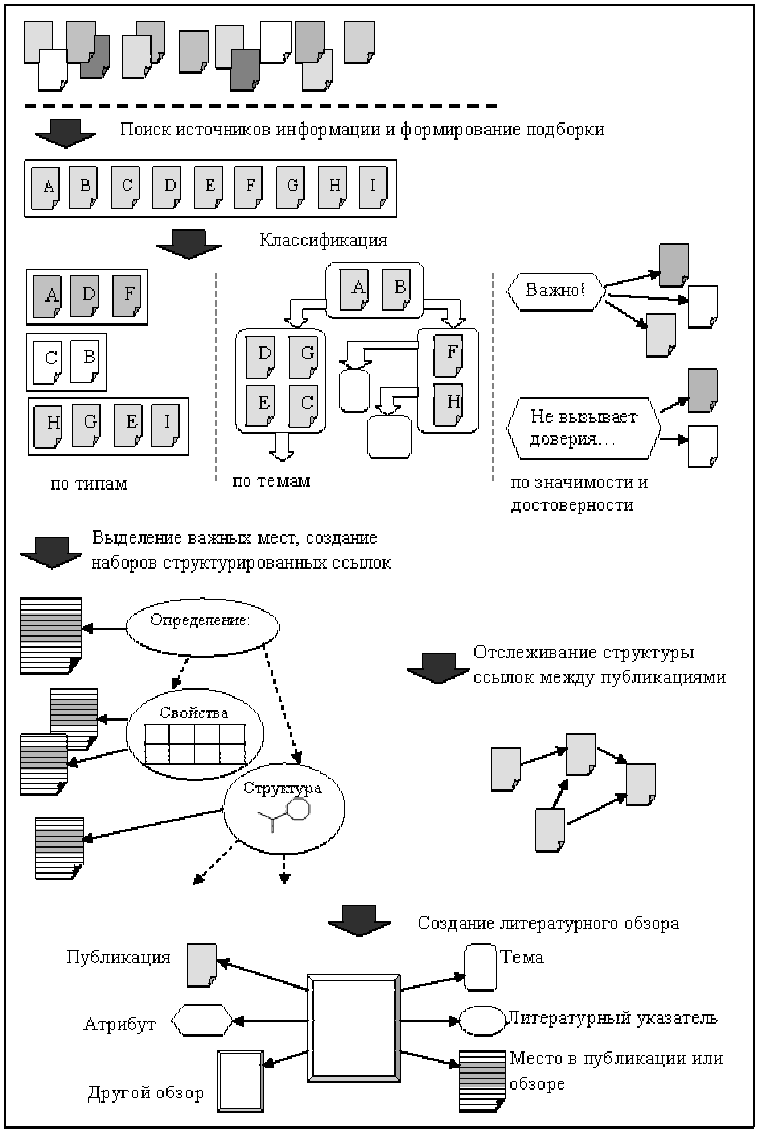

Рис. 1. Процесс обработки литературных данных

различных методик приготовления определённого катализатора или «отметок на полях» с комментариями к конкретным абзацам.);

• осуществление тематических выборок по теме, автору, организации и т. д.;

• хранение тематического обзора литературы на основе накопленных знаний. Обзор также может содержать множественные целевые ссылки на публикации и их части.

Предусмотрено также хранение произвольных структур данных, таких, как экспериментальные таблицы, фактографические и экспертные данные, путем создания в системе новых типов данных.

Работа с типами данных предполагает:

• добавление в систему новых структурных типов данных с возможностью наследования;

• семантическую классификацию типов.

Работа с данными предполагает:

- типизированную верификацию данных;
- каталогизацию данных одного типа;
- переиспользование данных;
- построение семантических иерархий объектов данных;
- установление ссылок из объектов пользовательских типов на различные объекты системы: публикации, обзоры, их темы и атрибуты и т. д.

Для решения проблемы создания произвольных типов данных, система предоставляет пользователю возможность описывать типы с использованием стандарта XML-схемы. Тип, созданный на основе XML-схемы, назван «схемо-типом», а экземпляр хранимого объекта схемо-типа – «объектом схемо-типа». Любые XML-документы, создаваемые на основе данных схем, проходят стандартную синтаксическую и структурную верификацию.

Наследование типов, определённых в схеме, а также установление ссылок между схемами позволяет эффективно определить множество типов, необходимых в данной экспертной области для описания специфических данных.

Для обеспечения возможности создания ссылок из одного XML-документа на другой система предоставляет возможность описывать подобные ссылки с помощью специальной системной схемы. В этом случае два XML-документа, построенные по двум семантически связанным схемам, могут ссылаться друг на друга, и это будет верифицироваться системой автоматически.

Для реализации описанной технологии была выбрана СУБД Oracle Release 10.2 ввиду расширенных возможностей работы с XML-данными.

Архитектура системы представлена на рис. 2.

Уровень данных – это уровень сервера базы данных. Здесь осуществляется хранение и проверка целостности данных.

Уровень данных состоит из трех частей. Внешний АПИ (API – application programming interface) – это программный интерфейс для клиентов, через который они могут осуществлять операции над объектами системы. Модель управления данными содержит основную логику обработки объектов системы. Схема данных – это набор таблиц и правил целостности данных системы.

Уровень сервера приложения представлен web-приложением, предоставляющим графический интерфейс пользователю, а также содержащий часть логики системы. Архитектура приложения выполнена на основе плагинов, что обеспечивает расширяемость системы за счёт новых плагинов. Центральными компонентами приложения являются Главный Контейнер и Менеджер Плагинов, главная цель которых состоит в создании среды для работы плагинов. Расширения представляют собой специальные модули приложения, которые позволяют плагинам легко приспособить существующую функциональность, как, например, отображение меню в верхней части web-страницы или обеспечения иерархического просмотра данных.

Наконец, плагины – это кирпичики системы, каждый из которых обеспечивает свою уникальную функциональность. Например, плагин Управления Схемо-Типами обеспечивает управление всеми схемо-типами системы.

Для представления разработанных интерфейсов доступа к XML-содержимому информационной системы через веб-интерфейс используется технология ASP.NET.

2. Создание тезаурусов и рубрикаторов по катализу

Решение всех перечисленных выше задач невозможно без создания словаря терминов предметной области, причем в этом словаре должны быть установлены связи между терминами и проведена классификация терминов. Следовательно, необходимо создание тезаурусов и рубрикаторов.

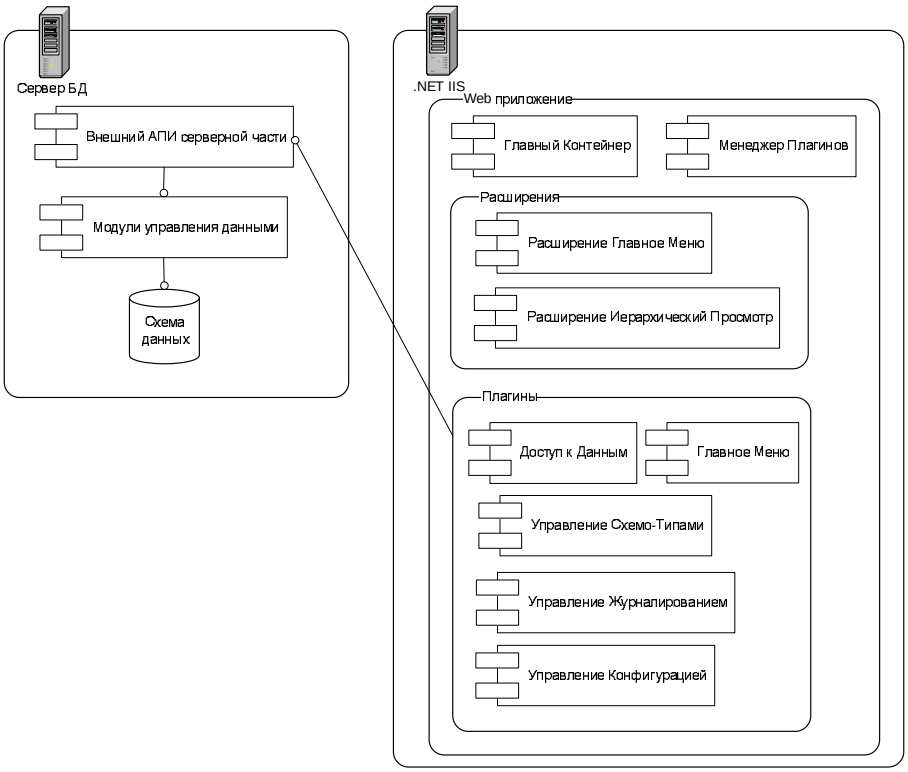

Рис. 2. Архитектура XML-системы хранения данных

Тезаурус – это максимально полный объем лексики, организованный по тематическому (семантическому) принципу с отражением определенного набора базовых семантических определений. Использование тезауруса в информационно-поисковых системах призвано повысить полноту поиска информации, позволяя расширять запрос синонимичными, более общими и более частными понятиями.

Рубрикатор – это систематизированный перечень наименований объектов, каждому из которых ставится в соответствие уникальный код. В настоящей работе рубрикатор рассматривается в качестве дерева тем или рубрик.

Для формального описания тезаурусов нами использована Zthes 1.0 XML-схема [4] и сопровождающее ее Дублинское Ядро [5]. Схема Zthes отвечает требованиям стандартов ISO 2788 (monolingual thesauri) и ISO 5964 (multilingual thesauri).

Для описания рубрикаторов разработана собственная описывающая их XML-схема (см. рис. 3).

Для решения стоящей перед нами задачи создания поисковой системы в области катализа в качестве исходного списка терминов и рубрик (а также для последующей работы по уточнению списков терминов и установлению связей) для тезаурусов и рубрикаторов были использованы предметный указатель и оглавление книг [6–9]. Данный подход был нами взят из работы [10], посвященной составлению тезауруса по предметной области «Математика».

При должном выборе предметный указатель вполне пригоден, если не в качестве полного, то, как минимум, в качестве базового списка ключевых слов, который при необходимости далее может пополняться.

В настоящее время создан тезаурус по гетерогенному катализу (около 7000 терминов) и два рубрикатора: по гетерогенному и промышленному гетерогенному катализу (примерно по 900 рубрик каждый).

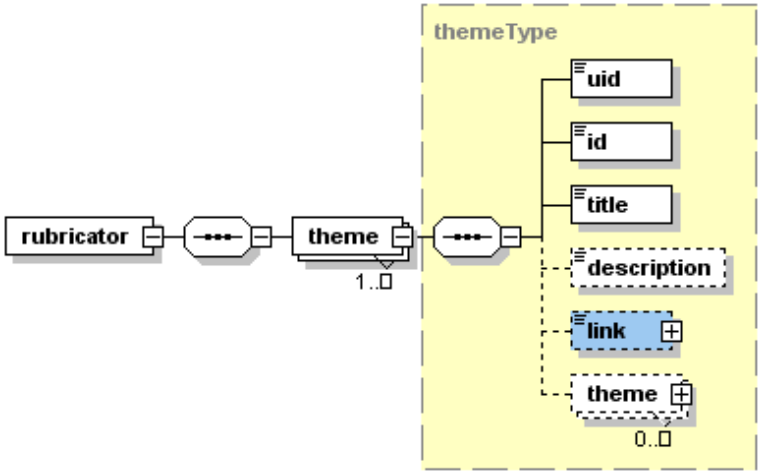

Рис. 3. XML-схема для описания рубрикатора

Между тезаурусами и рубрикаторами существует связь «многие-ко-многим», означающая, что термину из тезауруса может соответствовать несколько рубрик рубрикатора, а рубрике из рубрикатора – несколько терминов.

Например, термин «платина, катализаторы» связан с несколькими рубриками: «каталитическое гидрирование», «изомеризация алканов», «каталитическое окисление простых молекул», «окислительная конверсия метана в синтез-газ» и др. В свою очередь, рубрика «каталитическое гидрирование» связана с несколькими терминами: «ацетилен, селективное гидрирование», «бензол, гидрирование в циклогексан», «нитробензола гидрирование в анилин» и др. Для первоначальных версий тезауруса и рубрикатора такие связи были созданы автоматически путем установления связи между термином и рубрикой при наличии в предметном указателе для термина ссылки на страницу, относящуюся к данной рубрике (раздела в оглавлении).

Связь «многие-ко-многим» между терминами тезаурусов и рубриками рубрикаторов осуществляется посредством отдельного XML-файла, где для каждого термина указаны ссылки на рубрики. Данный подход хорош тем, что тезаурусы и рубрикаторы остаются независимыми объектами, но при этом отношение терминов к той или иной рубрике выясняется очень просто.

Для связывания терминов нами были использованы следующие типы связей, рекомендованные схемой Zthes 1.0.

1. NT (narrower term) – связанный термин имеет более узкое значение, чем текущий (связь с дочерним термином, то есть с термином более узкого смысла). Например, термин «метанол» с термином «метанол, окисление»;

2. BT (broader term) – связанный термин имеет более широкое значение, чем текущий (связь с родительским термином, то есть с термином более широкого смысла). Например, термин «метанол, окисление» с термином «окисление каталитическое». NT и BT взаимообратны;

3. USE (use instead) – связанный термин является дескриптором по отношению к текущему (связь с термином, который используется вместо этого), то есть связанный термин лучше употреблять, чем текущий, когда связанный термин более распространен или текущий термин – устаревший. Например, термины «метанол» и «древесный спирт» или в рассмотренном случае «молибдат железа, катализатор» и «железомолибдатный катализатор»;

4. UF (use for) – связанный термин является аскриптором по отношению к текущему (связь взаимообратная USE). Например, термины «древесный спирт» и «метанол»;

5. Связь RT (related term) – ассоциативная, то есть связь между терминами, которые нетождественны и которые трудно связать иерархически. Связь RT симметрична. Например, термины «метанол, окисление» и «кислород» напрямую не связаны, но при окислении метанола используют кислород;

6. Связь LE (linguistic equivalent) – связь между одинаковыми (тождественными) терминами на разных языках. Например, термины «метанол» и «methanol». Связь LE симметрична;

7. Дополнительная связь x-FE (full equivalent), которая обозначает полную тождественность терминов. Она симметрична. Данная связь введена нами как расширение схемы Zthes 1.0 в соответствии с рекомендациями разработчиков. Например, термины «формальдегид, окисление метанола», «метанол, окисление в формальдегид» и «формальдегид, получение окислением метанола».

Рассмотрим построение тезауруса на примере процесса производства формальдегида. Термин «формальдегид» является дочерним от термина «альдегиды», так как это – один из представителей этого класса химических соединений. Термин «формальдегид» связан с термином «формальдегид, производство» связью NT. Одним из способов производства формальдегида является окисление метанола. Поэтому от термина «формальдегид, производство» идет связь NT к термину «формальдегид, окисление метанола», который является тождественно связанным с термином «метанол, окисление в формальдегид» (связь FE). Термин «метанол, окисление в формальдегид», в свою очередь, связан с термином «метанол» связью BT и связью NT с термином «метанол, окисление, катализаторы». Процесс окисления метанола можно проводить с использованием в качестве катализаторов серебра, меди или молибдата железа. Соответственно, термины «серебро, катализатор окисления метанола», «медь, катализатор окисления метанола» и «молибдат железа, катализатор окисления метанола» являются дочерними по отношению к термину «метанол, окисление, катализаторы». В то же время медь, например, может быть не только катализатором окисления метанола, поэтому термин «медь, катализатор окисления метанола» является дочерним по отношению к термину «медь, катализатор», который, в свою очередь, связан с термином «медь» связью BT, поскольку медь может рассматриваться не только как катализатор. Рассмотренные катализаторы являются катализаторами окисления, следовательно, они связаны с термином «катализаторы окисления» связями BT. Термин «катализаторы окисления», в свою очередь, является дочерним термином по отношению к термину «катализаторы», так как «катализаторы» делятся на множество катализаторов других процессов. Процесс окисления метанола является примером процесса каталитического окисления, поэтому термин «окисление, каталитическое» является родительским по отношению к термину «метанол, окисление». Каталитическое окисление – это частный случай окисления. Окисление метанола осуществляется кислородом, но термин <кислород> только косвенно связан с терминами <окисление>, <окисление, каталитическое>, <метанол, окисление>, поэтому термин <кислород> имеет с ними ассоциативную связь RT. Термин «молибдат железа, катализатор» связан связью UF с термином «железомолибдатный катализатор», обратная связь между этими терминами – USE, поскольку термин «молибдат железа, катализатор» более широко применяется. Между терминами «кислород» и «oxygen» – связь LE.

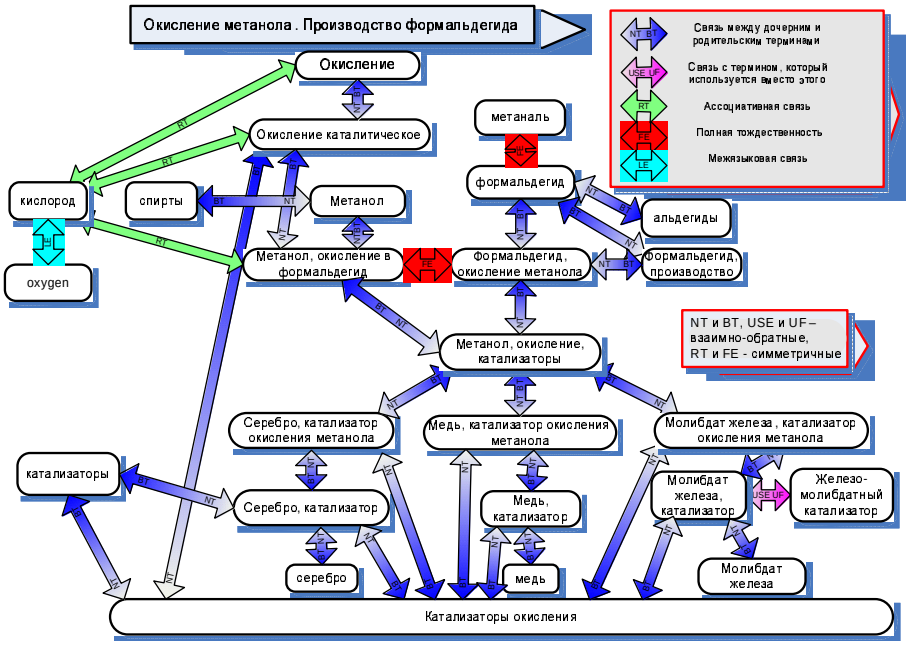

Рис. 4. Фрагмент тезауруса по гетерогенному катализу

При составлении тезауруса (см. рис. 4) используются следующие правила.

1. Если в название термина входит имя или фамилия автора, то название такого термина начинается с фамилии автора. Например, «Яна–Теллера эффект» или «Тёмкина изотерма адсорбции».

2. Названия терминов, состоящих их нескольких слов, образуются, начиная с наиболее значимого слова, но так, чтобы можно было понять, о чем идет речь. Например, «Теория кристаллического поля» записывается в термин «Кристаллического поля теория» или «Теория поля лигандов» – в «Поля лигандов теория».

3. Названия веществ выделяются в отдельные термины, а к ним в качестве дочерних относятся термины, обозначающие процессы, в которых используются эти вещества. Например, «пропилен» – BT по отношению к «пропилен, окисление в акролеин» или «платина, катализаторы» – BT по отношению к «платина, катализаторы гидрирования» и «платина, катализаторы углекислотной конверсии метана».

4. Термин, обозначающий процесс, является термином более широкого смысла, чем термин, обозначающий катализатор этого процесса. Они связаны связями NT и BT соответственно. Например, термин «родий, катализаторы парциального окисления метана» связан с термином «метан, парциальное окисление» связью BT или термин «железохромовый оксидный катализатор дегидрирования» связан с термином «этилбензол, дегидрирование в стирол» связью BT.

5. Между терминами на разных языках может быть только связь LE.

В настоящее время разработаны пользовательские интерфейсы, ориентированные на работу с тезаурусами и рубрикаторами в среде C#.NET. Идет работа по созданию веб-интерфейсов (см. рис. 5).

Необходимо отметить, что создание первых интернет-версий тезаурусов и рубрикаторов по катализу имеет важное значение безотносительно проблемы поиска в текстовых научных коллекциях.

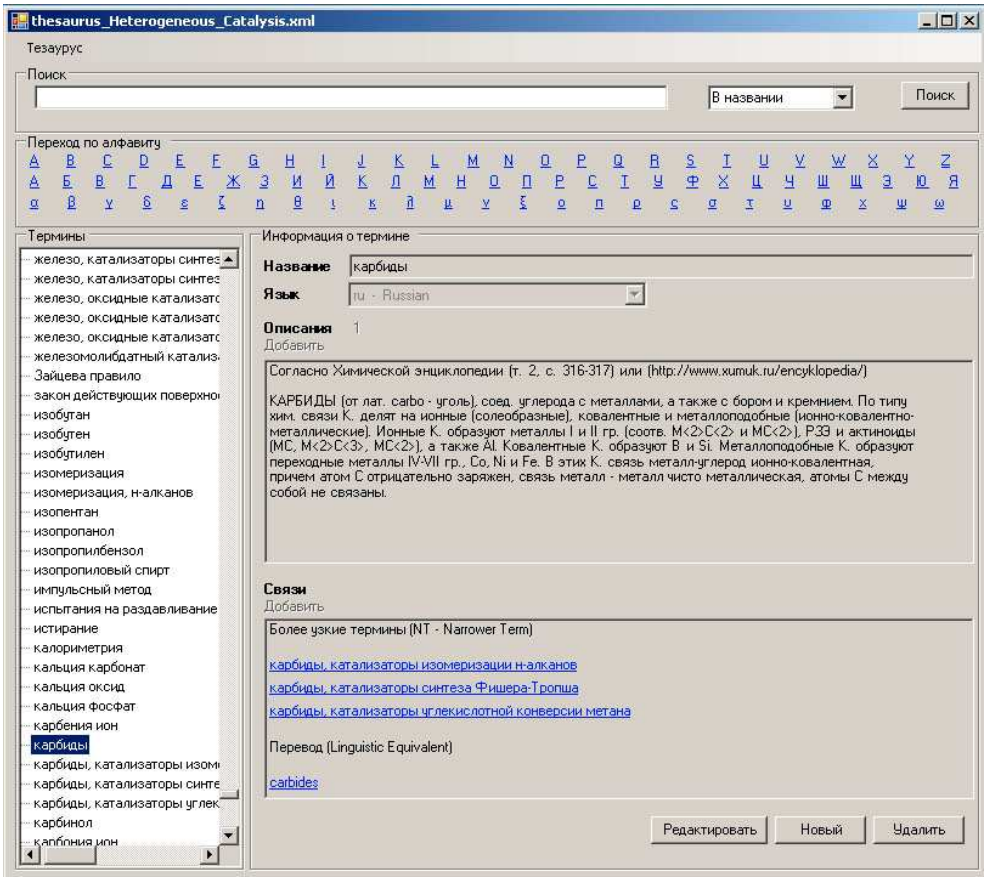

Рис. 5. Программное обеспечение для работы с тезаурусами

3. Создание поисковой системы по катализу с использованием сервиса Яндекс

Основной целью создания подсистемы поиска в научных текстовых источниках является существенное снижение требований к пользователю, касающихся его умения работать со сложными поисковыми запросами и наличия обширных знаний в предметной области (в данном случае – катализа). При этом поисковая система должна путем корректировки и дополнения (с использованием тезаурусов и рубрикаторов) запроса пользователя обеспечивать получение максимального числа релевантных результатов поиска.

Общая схема работы поиска такова: пользователь вводит первичный поисковый запрос для поискового модуля, который затем уточняется с помощью тезаурусов и рубрикаторов.

В качестве основы поискового модуля выбрано коммерческое программное обеспечение Yandex.Server [11]. Данный сервис может быть установлен на локальном сервере в нужной конфигурации и имеет множество достоинств: поиск словоформ, работа с учетом морфологии русского языка, модуль для работы с Oracle, работа с текстовыми данными произвольного формата (.pdf, .doc, xml и т. д.), вывод результатов в XML-формате. Кроме того, он позволяет решить проблему работы с базами данных большого объема ввиду использования для решения поисковых задач собственной индексной БД. Данный сервер обладает мощным

языком запросов (см. URL: http://company.yandex.ru/technology/products/YandexServer/searcher.xml#zapros). Кроме того, разработанный поисковый модуль может обращаться с запросом как к собственным текстовым базам данных, так и к поисковой системе Яндекс в сети Интернет.

После ввода пользователем первичного поискового запроса происходит его уточнение. Поисковый модуль обращается к тезаурусу, выбранному пользователем, который хранится в БД. Производится попытка выделить из запроса термины (планируется создать интерфейс поиска, в котором пользователь вводит термины в отдельные поля запроса, что позволяет избежать использования сложных анализаторов текста запроса) – каждое ключевое слово ищется в списке терминов указанного пользователем тезауруса. С учётом того, что каждый термин тезауруса может быть связан с одной или несколькими темами из рубрикатора, создается список тем, к которым могут относиться термины. Получившийся список предоставляется пользователю, из которого он (пользователь) выбирает нужную тему. Далее из тезауруса берутся термины, связанные с каждым найденным, причем только те, которые соотносятся с выбранной темой рубрикатора. Эти термины добавляются в запрос с учетом связей между ними и в соответствии с синтаксисом поискового сервера так, чтобы они одновременно расширяли и сужали запрос, создавая альтернативу и исключая ненужные термины.

Интерфейс подсистемы поиска в настоящий момент до конца не реализован, но предполагается, что его работа будет осуществляться в двух основных режимах – автоматическом и диалоговом.

Автоматическая часть:

1. В итоговый запрос добавляется исходный через логическое «ИЛИ».

2. Затем делается попытка разбить итоговый запрос на список терминов (CT). Для каждого найденного термина устанавливаются все его связи FE, USE, UF. Создаются запросы, где исходные термины заменены связанными, причем все слова должны искаться в границах одного предложения. Новые запросы добавляются в итоговый через логическое «ИЛИ».

3. Для каждого термина из CT берутся сужения (термины, связанные через NT), определяется пересечение. Если пересечение не пусто, то добавляются элементы пересечения в итоговый запрос через логическое «И».

Диалоговый режим:

1. Для терминов из CT просматриваются «глубокие сужения» – связи NT через один и более терминов, с согласия пользователя модифицируется запрос в соответствии с ними.

2. Для каждого термина из CT берутся сужения (термины, связанные через NT), попарно определяются пересечения. Пользователю предлагается изменить свой запрос в соответствии с элементами пересечений.

Сформированный запрос отправляется поисковому сервису Яндекс, который производит поиск и возвращает результаты в поисковый модуль.

Приведем пример работы с поисковой системой Яндекс в сети Интернет с использованием элементов данной методики. Для эксперимента взят исходный запрос «катализаторы производства формальдегида». При помощи только тезауруса запрос будет модифицироваться в соответствии с алгоритмом следующим образом:

1) к термину «формальдегид» будут найдены по связям FE термины «CH_2O» и «метаналь»;

2) далее в запрос будут добавлены пересечения сужений (NT) терминов «катализатор» и «производство формальдегида»: «серебро, катализатор», «медь, катализатор», «молибдат железа, катализатор»;

3) после приведения к языку запросов Яндекса и калибровки будет получен следующий запрос: (катализаторы &производства &(формальдегида | CH_2O | метаналь)) |(&катализаторы &производства &формальдегида && ((серебро & катализатор) | (медь & катализатор) | (молибдат &железа &катализатор))).

Проанализировав результаты обоих запросов, была подсчитана содержательная релевантность первых 10 и первых 20 результатов.

Для запроса «катализаторы производства формальдегида»:

- результат поиска: страниц – 3 336, сайтов – 564;
- релевантность: top20 = 70%, top10 = 60%.

Для итогового запроса:

- результат поиска: страниц – 1 303, сайтов – 139;
- релевантность: top20 = 85%, top10 = 100%.

Необходимо также отметить, что в будущем планируется расширить поисковую подсистему средствами классификации текстов. В числе основных методов, которые будут использованы для классификации текстов в соответствии с предпочтениями пользователя, можно назвать алгоритмы кластеризации документов на основании определения меры сходства новых документов с уже обработанными документами. Шкалами для вычисления меры сходства являются атрибуты описания документов. Для этой цели также используются предметные рубрикаторы и тезаурусы [12].

Заключение

Разработаны подходы к решению проблемы организации работы научного коллектива с хранилищем своих литературных и фактографических данных, а также к организации поиска в собственных текстовых коллекциях или в сети Интернет с использованием тезаурусов и рубрикаторов и поискового сервиса Яндекс. И хотя в настоящее время задача создания программного обеспечения полностью еще не завершена, предварительные полученные результаты дают уверенность в правильности выбранных подходов.

Работа поддержана Междисциплинарным Интеграционным проектом СО РАН № 111, «Фондом содействия отечественной науке», РФФИ (проект № 09-07-00298-а).

Summary

S.M. Khusainov, L.Yu. Ilyina, Anton O. Kuzmin, Andrey O. Kuzmin, V.N. Parmon. Specialized Factographic XML-Oriented Information System for Literary Data Storage and Search.

The article proposes an approach to solving the problem of organizing the work of a scientific team with storage of own literature data and factographic data. Organization of search within both own literary data archives and Internet is approached through the usage of thesauruses and rubricators, as well as Yandex search engine.

Key words: literary data proceeding, text searching, thesaurus on catalysis, rubricator on catalysis, science of catalysis.

Литература

1. *Нариньяни А.С.* ТЕОН-2: от Тезауруса к Онтологии и обратно // Труды междунар. семинара Диалог'2002 «Компьютерная лингвистика и интеллектуальные технологии». – М.: Наука, 2002. – Т. 1. – С. 199–154.

2. *Боровикова О.И., Загорулько Ю.А.* Организация порталов знаний на основе онтологий // Труды междунар. семинара Диалог'2002 «Компьютерная лингвистика и интеллектуальные технологии». – Протвино, 2002. – Т. 2. – С. 76–82.

3. *Загорулько Ю.А., Боровикова О.И., Кононенко И.С., Сидорова Е.А.* Подход к построению предметной онтологии для портала знаний по компьютерной лингвистике // Компьютерная лингвистика и интеллектуальные технологии: Труды междунар. конф. «Диалог 2006» (Бекасово, 31 мая – 4 июня 2006 г.). – М.: Изд-во РГГУ, 2006. – С. 148–151.

4. Zthes XML Schema v. 1.0. – URL: http://zthes.z3950.org/schema/index.html.

5. Dublin Core. – URL: http://purl.org/dc/elements/1.1/.

6. Handbook of Heterogeneous Catalysis: In 5 V. / Eds. G. Ertl, H. Knozinger, J. Weitkamp. – Weinheim: Wiley-VCH, 1997.

7. *Крылов О.В.* Гетерогенный катализ. – М.: Академкнига, 2004. – 679 с.

8. *Сеттерфильд Ч.* Практический курс гетерогенного катализа. – М.: Мир, 1984. – 521 с.

9. Химическая энциклопедия: в 5 т. – М.: Сов. энцикл., 1988.

10. *Барахнин В.Б.* Разработка тезауруса предметной области «Математика» // Вычислительные технологии, Т. 8; Региональный вестник Востока, № 3 (19): Совместный вып. – 2003. – Ч. 1. С. 111–115.

11. Яndex.Server. – URL: http://company.yandex.ru/technology/products/yandex-server.xml.

12. *Барахнин В.Б., Нехаева В.А., Федотов А.М.* Методика отбора публикаций из библиографических баз данных на основании меры сходства // Материалы Всерос. конф. с междунар. участ. «Знания – Онтологии – Теории» (ЗОНТ-07) (Новосибирск, 14–16 сент. 2007 г.). – Новосибирск, 2007. – Т. 2. – С. 88–94.

Поступила в редакцию
26.02.09

Хусаинов Сергей Муратович – аспирант Института систем информатики СО РАН, г. Новосибирск.

Ильина Людмила Юрьевна – кандидат химических наук, научный сотрудник Института катализа им. Г.К. Борескова СО РАН, г. Новосибирск.
E-mail: *ilud@catalysis.ru*

Кузьмин Антон Олегович – кандидат физико-математических наук, ведущий программист Института катализа им. Г.К. Борескова СО РАН, г. Новосибирск.
E-mail: *ao_kuzmin@list.ru*

Кузьмин Андрей Олегович – кандидат химических наук, старший научный сотрудник Института катализа им. Г.К. Борескова СО РАН, г. Новосибирск, старший преподаватель Новосибирского государственного университета.
E-mail: *kuzmin@catalysis.ru*

Пармон Валентин Николаевич – академик РАН, директор Института катализа им. Г.К. Борескова СО РАН, г. Новосибирск.
E-mail: *parmon@catalysis.ru*