ТЕХНОЛОГИЯ СОЗДАНИЯ ТЕЗАУРУСА ПРЕДМЕТНОЙ ОБЛАСТИ НА ОСНОВЕ ПРЕДМЕТНОГО УКАЗАТЕЛЯ ЭНЦИКЛОПЕДИИ

В. Б. Барахнин
Институт вычислительных технологий СО РАН, Новосибирск, Россия
e-mail: bar@ict.nsc.ru

В. А. Нехаева
Новосибирский государственный университет, Россия
e-mail: nekhaeva@ngs.ru

This work describes a technology for creation of object domain thesaurus, which is based on subject heading for specialized encyclopedia. Such technology offers a high quality description of the object domain using reliable terms thus allowing to build up a first stage of thesaurus with a minimal engagement of experts in this particular field of knowledge. The proposed technology also contains a thesaurus building algorithm and web based application implementing this algorithm.

Введение

Одним из важнейших факторов, обеспечивающих успешное выполнение интеграционных научно-исследовательских проектов, является эффективное научно-информационное обеспечение. В частности, совместная работа исследователей нескольких (притом не всегда смежных) специальностей требует тщательного согласования используемой терминологии, ибо одно и то же понятие может обозначаться в разных областях науки различными терминами, а одним термином — разные понятия.

Другая задача информационного обеспечения проектов — создание интегрированной картотеки библиографических описаний документов (т. е. статей, книг и т. д.) по тематике проекта, составленной путем объединения ресурсов совместно работающих исследователей, у каждого из которых за годы его работы уже накоплена картотека по той или иной тематике (в настоящее время подобные картотеки хранятся, как правило, на электронных носителях). Для облегчения поиска в картотеке желательно, чтобы ключевые слова, характеризующие документы, выбирались по возможности из единого словаря. Для автоматической классификации документов, включенных в картотеку или потенциально могущих быть занесенными в нее из электронных баз данных

научных публикаций типа базы данных реферативных журналов, “Current Contents” и т. п., представляется целесообразным использовать алгоритм координатного индексирования [1]. Этот алгоритм основан на учете классификационных признаков входящих в текст терминов (слов и словосочетаний), характеризующих ту или иную предметную область.

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. Такой словарь называется тезаурусом (см. подробнее в [1]). Тезаурус (или нормативный тезаурус) — это словарь-справочник, содержащий все лексические единицы информационно-поискового языка — дескрипторы (вместе с ключевыми словами, которые в пределах данной информационно-поисковой системы считаются синонимами этих дескрипторов), причем дескрипторы в словаре должны быть систематизированы по смыслу, а смысловые связи между ними эксплицитно выражены.

Однако составление тезауруса “с чистого листа” может потребовать весьма значительных трудозатрат специалистов-экспертов, которые должны собрать все термины, достаточно полно охватывающие предметную область, согласовать их значения, установить связи и провести классификацию. Подобные трудности, возникающие при решении хотя и важной, но все-таки вспомогательной задачи, способны негативно повлиять на перспективы ее решения.

Нами разработана и реализована технология создания тезауруса на основе предметного указателя специализированных энциклопедий. Эта технология обеспечивает высококвалифицированное описание предметной области с использованием надежно выверенных терминов, позволяя провести начальный этап построения тезауруса с минимальным привлечением специалистов — экспертов в данной предметной области. Подробное изложение и обоснование алгоритма даны в работе [2]. Ниже приведено краткое описание алгоритма, а также реализующего его web-приложения.

1. Алгоритм создания тезауруса

В качестве списка ключевых слов и словосочетаний для тезауруса предлагается использовать предметный указатель специализированной энциклопедии (или нескольких энциклопедий). Выбор конкретной энциклопедии осуществляет специалист по предметной области, и этот выбор зависит от целей, преследуемых при создании тезауруса. Так, для решения комплексных экологических задач целесообразно использовать энциклопедии (или, при их отсутствии, — энциклопедические словари) по физике, химии, геологии, биологии, медицине, математике и т. п. При должном выборе предметный указатель вполне пригоден если не в качестве полного, то, как минимум, в качестве базового списка ключевых слов, который при необходимости будет пополняться.

Предметные указатели большей части энциклопедий устроены сходным образом — в них содержатся термины, являющиеся названиями статей энциклопедии, термины, определения которых даны в статьях, а также упомянутые в статьях наиболее важные результаты.

В качестве дескрипторов (т. е. терминов, являющихся именами классов близких по смыслу понятий) полагаются названия статей энциклопедии, а связанными с ними по смыслу считаются слова из предметного указателя, встречающиеся в соответствующих

статьях. Основным преимуществом такого метода является то, что для установки типов связей между терминами не требуется быть экспертом в данной предметной области — вполне хватит общих знаний, позволяющих понять текст энциклопедии, — более конкретные сведения, необходимые в процессе классификации понятий, всегда можно почерпнуть из конкретной статьи.

Поскольку создаваемый тезаурус предназначен для работы с использованием протокола Z39.50, типы связей устанавливаются в соответствии с рекомендациями схемы Zthes [3], которая выделяет следующие типы:

— **BT** — связь с родительским термином, т. е. с термином более широкого смысла;

— **NT** — связь с дочерним термином, т. е. с термином более узкого смысла. Связь **BT** — **NT** является взаимно-обратной;

— **USE** — связь с термином, который используется вместо этого;

— **UF** — взаимно-обратная связь **USE**;

— **RT** — связь, определяющая связанный по смыслу термин;

— **LE** — связь между лингвистически эквивалентными терминами;

— **FE** — полностью тождественные термины.

Далее проводится классификация дескрипторов в соответствии с разделами данной предметной области. Выбор конкретного классификатора, как и выбор энциклопедии, осуществляется специалистом-экспертом, причем в случае использования нескольких энциклопедий из разных предметных областей возможно использование нескольких специализированных классификаторов. Между дескрипторами и разделами классификатора устанавливаются связи вида **NT**, **RT**, **LE** (**FE**), при этом при классификации следует использовать, по возможности, разделы максимально низкого уровня.

После этого ключевым словам, связанным с дескриптором отношениями **BT**, **USE**, **RT**, **LE** и **FE**, приписывается тот же классификационный номер, что и дескриптору. Впрочем, это не исключает такой ситуации, что если дескриптор отнесен к классу не самого низкого уровня, то при последующей работе эксперта термины, связанные с дескриптором отношениями **BT** и **USE**, могут быть отнесены к классу более низкого уровня. В этом случае указанные термины сами станут дескрипторами.

В результате все термины, входящие в предметный указатель, оказываются рассклассифицированы в соответствии с разделами данной предметной области.

2. Описание работы web-приложения

Тем не менее процесс построения тезауруса в соответствии с данной методикой подразумевает большой объем рутинной работы и, кроме того, требует участия человека, имеющего навыки программирования. Поэтому в дополнение к методике было разработано web-приложение, обладающее дружественным к пользователю интерфейсом и поддерживающее следующие функции:

1) автоматический перевод информации с оцифрованных страниц предметного указателя в таблицу базы данных;

2) выделение дескрипторов в общем списке терминов;

3) поиск терминов, связанных с данным дескриптором, и установка типов связей в соответствии со схемой Zthes.

Важно отметить, что для выполнения всех упомянутых выше операций навыков программиста не требуется.

Разработанное приложение является универсальным, т. е. может быть использовано для создания тезаурусов различных предметных областей. В настоящий момент перенастройку программы с предметного указателя одной энциклопедии на предметный указатель другой (а лишь на этом этапе процессы построения тезаурусов разных предметных областей могут различаться) выполняет программист, однако ведутся работы по дополнению программы функциями, позволяющими проводить эту операцию пользователю, не имеющему навыков программирования.

Функционирует приложение следующим образом. Обработка оцифрованных страниц предметного указателя производится автоматически. Пользователь указывает местоположение текстового файла с данными, после чего происходит его построчное считывание и в базу данных заносятся сами термины, а также информация о номерах страниц энциклопедии, где они расположены (рис. 1).

Дескрипторы из общего списка ключевых слов выделяет сам пользователь, отмечая искомые термины в выведенном на экран списке. Web-приложение поддерживает также функцию исправления возможных ошибок (рис. 2). Напомним, что связанными с данным дескриптором считаются все термины, встречающиеся в посвященной ему статье энциклопедии.

Для облегчения поиска связанных терминов пользователю выводится только список ключевых слов, расположенных на той же странице, что и выбранный им дескриптор (собственно, для этого мы и заносили в базу данных не только термины, но и информацию о номерах страниц). Разумеется, поскольку статья может занимать не всю страницу целиком, в список попадут лишние термины. Пользователь, устанавливая связи,

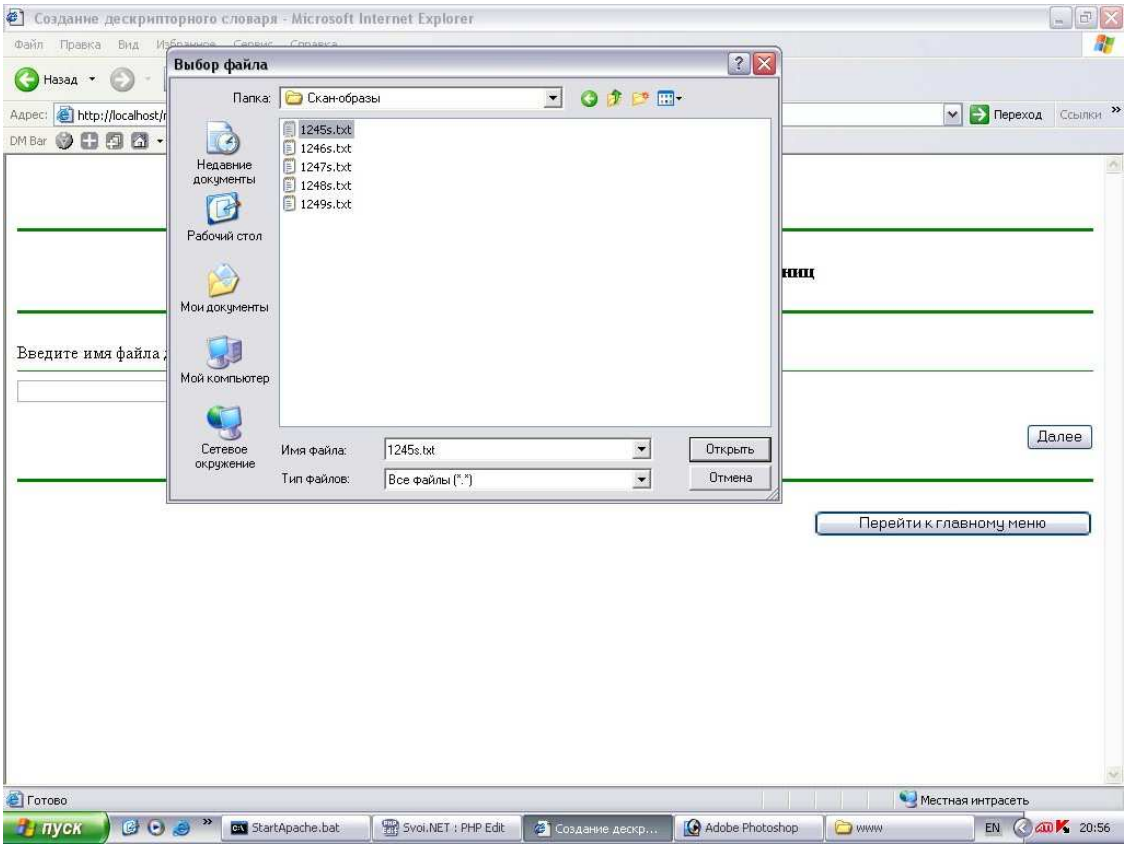

Рис. 1. Занесение текстовых файлов с терминами из предметного указателя

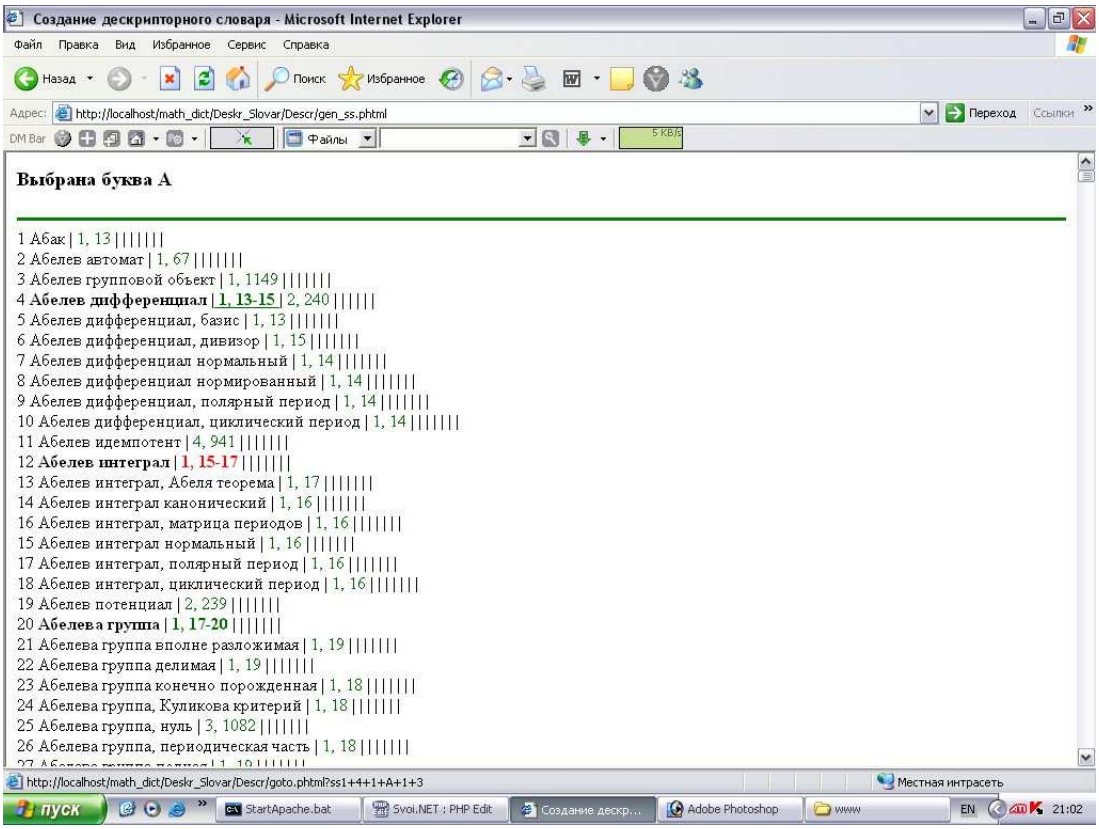

Рис. 2. Список ключевых слов и выделение дескрипторов

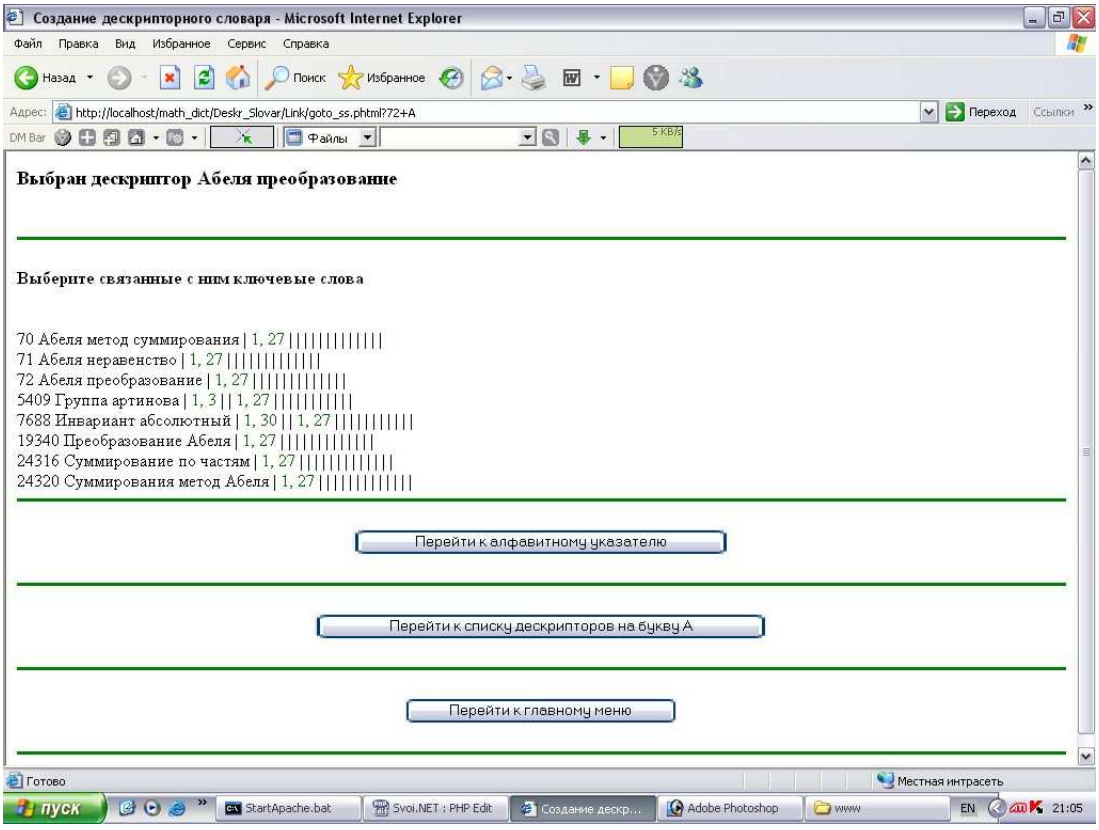

Рис. 3. Выбор связанных терминов

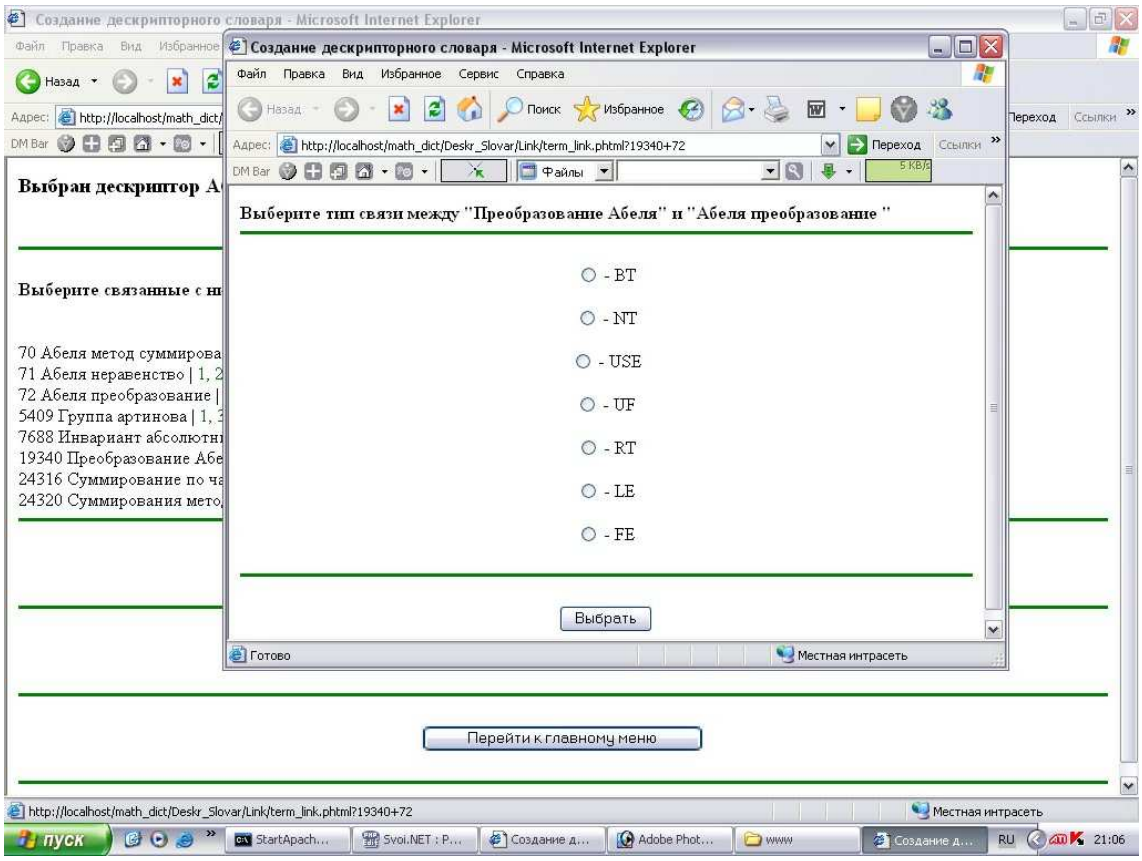

Рис. 4. Установление типов связей.

выберет лишь часть ключевых слов из предложенного списка, однако и такая автоматизация заметно снижает объем рутинной работы (рис. 3).

Тип связи между дескриптором и ключевым словом уточняется путем заполнения соответствующей формы (рис. 4).

Заключение

Работоспособность данного алгоритма и web-приложения была проверена путем создания тезауруса ряда разделов предметной области “Математика” (“Дифференциальные уравнения”, “Уравнения в частных производных”, “Численный анализ”, “Механика жидкости” и др.) на основе предметного указателя “Математической энциклопедии”. Установлено, что для классификации терминов и установления связей между ними достаточно квалификации бакалавра (при условии привлечения в редких случаях для консультаций эксперта с ученой степенью). Это доказывает высокую эффективность разработанного алгоритма.

Список литературы

[1] Михайлов А.И., Черный А.И., Гиляревский Р.С. Основы информатики. М.: Наука, 1968.

[2] Барахнин В.Б. Разработка тезауруса предметной области “Математика” // Матер. конф. “Вычислительные и информационные технологии в науке, технике и образовании”. Ч. 1. Новосибирск; Алматы; Усть-Каменогорск, 2003. С. 111–115.

[3] Zthes: a Z39.50 Profile for Thesaurus Navigation
`http://lcweb.loc.gov/z3950/agency/profiles/zthes-04.html`

Поступила в редакцию 11 мая 2007 г.