АВТОМАТИЗИРОВАННОЕ СОЗДАНИЕ ТЕЗАУРУСА ТЕРМИНОВ ПРЕДМЕТНОЙ ОБЛАСТИ ДЛЯ ЛОКАЛЬНЫХ ПОИСКОВЫХ СИСТЕМ

Виталий Величко, Павел Волошин, Светлана Свитла

***Аннотация**: В статье рассмотрен метод автоматизированного создания тезауруса терминов предметной области на основе синтактико-семантического анализа естественно-языковых текстов для повышения релевантности поиска в полнотекстовых локальных поисковых системах. Использование предложенного метода позволяет сократить затраты времени на составление и редактирование тезауруса.*

***Ключевые слова**: локальные полнотекстовые поисковые системы, тезаурус терминов, синтактико-семантический анализ.*

***ACM Classification Keywords**: I.2.7 Natural Language Processing - Text analysis*

***Conference**: The paper is selected from XVth International Conference "Knowledge-Dialogue-Solution" KDS-2 2009, Kyiv, Ukraine, October, 2009.*

Введение

Количество электронных документов, которые использует в своей ежедневной деятельности современная компания, стремительно возрастает. При этом данные хранятся в различных хранилищах, каждое из которых имеет собственную структуру (базы данных, информационные порталы, электронные библиотеки и т.д.) либо хранилище документов вообще неструктурированно (файлы на жестком диске пользователя).

Поэтому для обеспечения жизнедеятельности крупных государственных структур и частных корпораций необходимым условием является использование локальных поисковых систем для осуществления поиска по внутренним информационным ресурсам.

Одними из основных требований к подобным системам являются:

- обязательная полнотекстовая индексация всех информационных ресурсов, в которых осуществляется поиск, независимо от типов файлов и структуры хранения данных;
- наличие лингвистического процессора для выделения лексем, который позволяет осуществлять поиск по всем падежным формам искомого слова или словосочетания, что особенно важно для флективных языков, в частности, русского и украинского языка;
- упорядочивание результатов поиска по релевантности найденных документов.

На сегодняшний день в Украине наибольшее распространение получили локальные поисковые системы, такие как META, Google Desktop Search, Yandex.Server. В тоже время, все большее распространение и популярность приобретает программный продукт компании Microsoft — "Microsoft Office SharePoint Server 2007", который является средством организации коллективной работы с корпоративными данными и используется для решения чрезвычайно широкого спектра задач [1]:

- управление информационным порталом и бизнес-данными;
- поиск по данным предприятия;
- коллективная работа с документами;

- бизнес-аналитика;
- автоматизация деловых процессов.

Поиск по данным предприятия, осуществляемый Microsoft Office SharePoint Server 2007, не только удовлетворяет всем вышеперечисленным основным требованиям к локальным поисковым системам, но и имеет такие преимущества как: легкость администрирования; настройка страницы вывода результатов поиска; поиск не только по содержимому, но и по свойствам документа; возможность подключения списков расширения и замещения для управления поисковым запросом и т.д.

Список замещения определяет шаблон, который заменяется в поисковом запросе одним или несколькими словами – подстановками. Например, "*рост*" шаблон, а "*возрастание*" - подстановка. При вводе поискового запроса "*рост*" поисковый механизм отобразит результаты поиска только для поискового запроса "*возрастание*". Результаты поиска для слова "рост" отображаться не будут. Список расширения – это группа подстановок, которые являются синонимами. Запросы, содержащие слова из множества подстановок, расширяются путем включения всех других подстановок. Например, можно создать список расширения, где следующие подстановки являются синонимами: "*возрастание*", "*наращивание*", "*расширение*". При вводе поискового запроса "*возрастание*" поисковый механизм отобразит результаты поиска также для слов "*наращивание*" и "*расширение*".

Однако, все вышеперечисленные поисковые системы не имеют встроенных списков терминов (тезаурусов), учитывающих синонимию понятий предметной области. Для получения наиболее полных результатов поиска, пользователям в поисковых запросов приходится постоянно перечислять все синонимы введенного термина (например, *Россия* – это и *Российская Федерация*, и *РФ*).

При наличии тезауруса терминов предметной области, пользователю в поисковом запросе достаточно ввести только один термин. Если в тезаурусе есть список синонимов к введенному слову, то в результатах поиска будут присутствовать как документы, которые содержат слово, введенное пользователем, так и документы, содержащие слова-синонимы.

К сожалению, из-за отсутствия формализованных словарей терминов для конкретных предметных областей, автоматическое создание тезауруса невозможно. Ручное составление тезауруса является весьма трудоемкой задачей, так как требует экспертного анализа значительного количества документов организации (корпорации) для выделения списка терминов предметной области, при этом достаточно трудно оценить полноту полученного списка. Для решения такой задачи необходимо использовать автоматизированное создание списка терминов предметной области.

Принципы автоматического выбора терминов

Для построения понятийного аппарата из текстов предметной области используется поиск и выделение субстантивных именных словосочетаний, выражаемых схемой: согласуемое слово + существительное. В этой модели существительное является главным словом, а согласуемое слово — зависимым и может выражаться как прилагательным, так и существительным [2]. Словосочетания могут включать в свой состав также предлоги и сочинительные союзы. Количество слов в именных словосочетаниях колеблется от двух до пятнадцати и в среднем составляет три слова [3]. В работе [2] приводится 18 шаблонов именных словосочетаний, используемых для выделения терминов предметной области. В русском языке синтаксическая структура терминов предметной области более чем в 90 процентов случаев соответствует следующим пяти шаблонам: одиночные существительные, прилагательные, и сокращения; существительное + существительное в родительном падеже; прилагательное + существительное;

прилагательное + прилагательное + существительное; существительное + прилагательное + существительное в родительном падеже [4].

Вместе с тем существуют сложные словосочетания, используемые для обозначения понятий и терминов, состоящих из трех и более значимых слов. Выражение понятий и терминов словосочетаниями в пять и более слов, с использованием союзов и предлогов встречается редко, особенно такими словосочетаниями, в которых части речи не чередуются (например, прилагательное + прилагательное + прилагательное + существительное + существительное в родительном падеже). Эксперименты по выделению терминов показали, что в украинском языке для предметной области "экономика" целесообразно увеличить количество слов в синтаксической структуре именных словосочетаний до пяти. Словосочетания длиной пять и более слов используются в наименованиях организаций, в определении понятий относящихся к финансово-экономической сфере деятельности организаций. Шаблоны именных словосочетаний, используемых для поиска терминов, приведены в Таблице 1.

Таблица 1

№	Структура шаблона	Пример термина
1	Аббревиатура	СНГ; МВФ
2	Существительное	Система; государство
3	Существительное + существительное_в_родительном_падеже	Президент Украины; Глава государства
4	Прилагательное + существительное	Конституционный Суд; экономический рост
5	Существительное + существительное_в_родительном_падеже + существительное_в_родительном_падеже	Министр финансов Украины; Указ Президента Украины
6	Прилагательное + существительное + существительное_в_родительном_падеже	Финансовая система государства; Верховная Рада Украины; Бюджетный кодекс Украины
7	Существительное + прилагательное_в_родительном_падеже + существительное_в_родительном_падеже	Органы исполнительной власти; девальвация национальной валюты
8	Прилагательное + прилагательное + существительное	Среднемесячная заработная плата
9	Существительное + существительное_в_родительном_падеже + существительное_в_родительном_падеже + существительное_в_родительном_падеже	Постановление Кабинета Министров Украины; архив Министерства обороны Украины
10	Прилагательное + существительное + существительное_в_родительном_падеже + существительное_в_родительном_падеже	Национальная программа иммунопрофилактики населения
11	Существительное + прилагательное_в_родительном_падеже + существительное_в_родительном_падеже + существительное_в_родительном_падеже	Средства Государственного бюджета Украины; Председатель Верховной Рады Украины

№	Структура шаблона	Пример термина
12	Прилагательное + прилагательное + существительное + существительное_в_родительном_падеже	Государственная налоговая администрация Украины
13	Существительное + существительное_в_родительном_падеже + прилагательное_в_родительном_падеже + существительное_в_родительном_падеже	Система использования бюджетных средств; орган управления государственной службой
14	Прилагательное + существительное + прилагательное_в_родительном_падеже + существительное_в_родительном_падеже	Центральные органы исполнительной власти
15	Существительное + прилагательное_в_родительном_падеже + прилагательное_в_родительном_падеже + существительное_в_родительном_падеже	Зона чрезвычайной экологической ситуации; последствия мирового финансового кризиса
16	Прилагательное + прилагательное + прилагательное + существительное	Независимый государственный финансовый контроль
17	Прилагательное + существительное + существительное_в_родительном_падеже + существительное_в_родительном_падеже + существительное_в_родительном_падеже	Финансово-экономическая деятельность Министерства обороны Украины
18	Существительное + прилагательное_в_родительном_падеже + существительное_в_родительном_падеже + существительное_в_родительном_падеже + существительное_в_родительном_падеже	Комитет финансового контроля Министерства финансов
19	Прилагательное + прилагательное + существительное + существительное_в_родительном_падеже + существительное_в_родительном_падеже	Обесцененные денежные сбережения граждан Украины
20	Существительное + существительное_в_родительном_падеже + прилагательное_в_родительном_падеже + существительное_в_родительном_падеже + существительное_в_родительном_падеже	Фонд содействия молодежному жилищному строительству; нормы бюджетного законодательства Украины
21	Прилагательное + существительное + прилагательное_в_родительном_падеже + существительное_в_родительном_падеже + существительное_в_родительном_падеже	Специальный фонд государственного бюджета Украины
22	Существительное + прилагательное_в_родительном_падеже + прилагательное_в_родительном_падеже + существительное_в_родительном_падеже + существительное_в_родительном_падеже	Управление государственными финансовыми ресурсами Украины

№	Структура шаблона	Пример термина
23	Существительное + существительное_в_родительном_падеже + существительное_в_родительном_падеже + прилагательное_в_родительном_падеже + существительное_в_родительном_падеже	Источник финансирования дефицита Государственного бюджета; реформа системы контроля государственных финансов
24	Прилагательное + прилагательное + прилагательное + существительное + существительное_в_родительном_падеже	Государственный внебюджетный Пенсионный фонд Украины
25	Прилагательное + существительное + существительное_в_родительном_падеже + прилагательное_в_родительном_падеже + существительное_в_родительном_падеже	Главный распорядитель средств Государственного бюджета
26	Существительное + прилагательное_в_родительном_падеже + существительное_в_родительном_падеже + прилагательное_в_родительном_падеже + существительное_в_родительном_падеже	Контроль экономической деятельности коммерческих организаций
27	Прилагательное + прилагательное + существительное + прилагательное_в_родительном_падеже + существительное_в_родительном_падеже	Международный аудиторский контроль финансовой деятельности
28	Существительное + существительное_в_родительном_падеже + прилагательное_в_родительном_падеже + прилагательное_в_родительном_падеже + существительное_в_родительном_падеже	Система управления государственными финансовыми ресурсами
29	Прилагательное + существительное + прилагательное_в_родительном_падеже + прилагательное_в_родительном_падеже + существительное_в_родительном_падеже	Государственное регулирование внешней экономической деятельности
30	Существительное + прилагательное_в_родительном_падеже + прилагательное_в_родительном_падеже + прилагательное_в_родительном_падеже + существительное_в_родительном_падеже	Орган внешнего государственного финансового контроля
31	Существительное + существительное_в_родительном_падеже + существительное_в_родительном_падеже + существительное_в_родительном_падеже + существительное_в_родительном_падеже	Программа деятельности Кабинета Министров Украины

Автоматическое выделение однословных и многословных терминов, кроме шаблонов, использует результаты синтактико-семантического анализа текста. Распознание поверхностных семантических отношений осуществляется с помощью анализа флексий полнозначных слов, учитывая предлоги и союзы, без предварительного полного грамматического разбора и построения синтаксических отношений, которые используется в традиционной грамматике [5].

Процедура выделения терминов из текста включает два основных этапа.

На первом этапе происходит непосредственный поиск в тексте слов и словосочетаний – кандидатов в термины. В качестве однословных терминов выбираются существительные и аббревиатуры. Многословные термины формируются с помощью определенных типов отношений между словами предложения, путем постепенного присоединения слов к однословному термину-существительному. Для терминов – именных словосочетаний используются следующие основные типы отношений между словами: *объектное*, *принадлежность* (между двумя существительными), *определительное* (между прилагательным и существительным), *однородные слова* (между двумя существительными или двумя прилагательными). Выделенные группы слов проверяются на соответствие заданным шаблонам, приведенным в Таблице 1. Порядок расположения в предложении слов, образующих термин, может точно не соответствовать заданному шаблону, но обязательным условием выделения термина является соответствие отношений между словами определенным типам отношений. Это позволяет, например, из предложения "*Построение онтологии указанной предметной области*" выделить термин "*онтология предметной области*".

На втором этапе список кандидатов в термины фильтруется: учитывается значимость выделенных словосочетаний (приближенность в дереве разбора к подлежащему или сказуемому предложения) и частота, с которой они встречаются в тексте.

Приведенные принципы автоматического построения списка возможных терминов реализованы для украинского и русского языков в программе "Конспект".

Построение тезауруса терминов

Полученный предварительный список терминов редактируется вручную с помощью утилиты – редактора тезауруса терминов предметной области (общий вид окна редактора изображен на Рис. 1.)

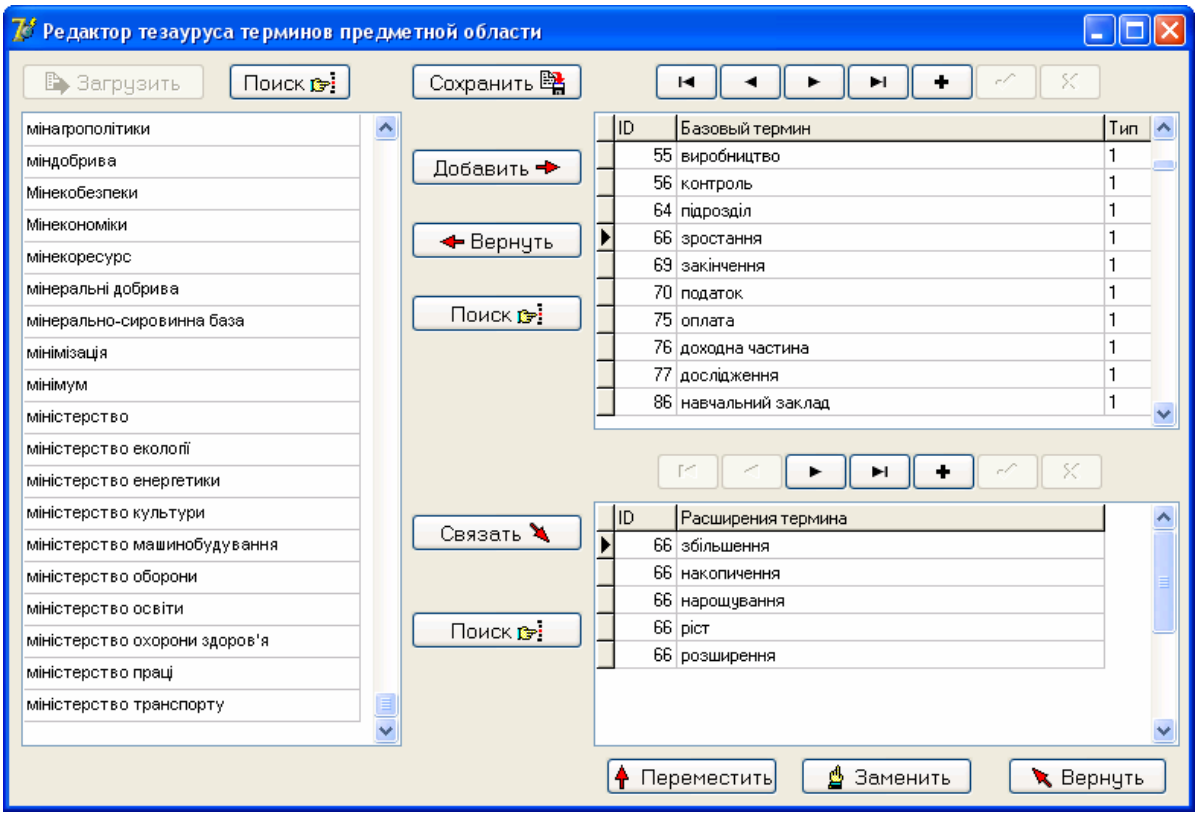

Рисунок 1. Главная форма редактора тезауруса терминов предметной области

Входными данными для утилиты является список терминов, сформированный программой "Конспект". Эксперт-аналитик вручную связывает термины, являющиеся синонимами для заданной предметной области (см. Рис.1). Полученные кортежи синонимов терминов предметной области сохраняются в XML-файл заданной структуры, который может использоваться поисковой системой среды Microsoft Office SharePoint Server 2007 в качестве тезауруса (списка расширений). В общем виде процесс автоматизированного построения тезауруса терминов предметной области изображен на Рис. 2.

Рисунок 2. Схема процесса автоматизированного построения тезауруса терминов

Эксперименты

Рассмотренный метод автоматизированного создания тезауруса терминов предметной области был использован для обработки текстов на украинском языке, относящихся к финансово-экономической сфере деятельности организаций. Из 172 различных текстовых документов за 2000-2008 года общим объемом примерно 3000 страниц текста без форматирования (примерно 12 Мб файлов формата txt) автоматически было выделено 26 860 слов и словосочетаний. Из сформированного списка для дальнейшего ручного редактирования терминов было оставлено 6,5 тыс. слов и словосочетаний, которые встречались в исходных текстов более трех раз. Термины, не имеющие синонимов, были исключены из тезауруса. В качестве базовых терминов было выбрано около 400 слов и словосочетаний, а в список расширений добавлено примерно 650 терминов. От общего количества терминов в тезаурусе однословные термины составили 76%, двухсловные – 21%, термины, состоящие из трех и более слов– 3%.

Выводы

В статье рассмотрена методика автоматизированного создания тезауруса терминов предметной области на основе синтактико-семантического анализа естественно-языковых текстов. Приведенные принципы автоматического построения списка возможных терминов реализованы для украинского и русского языков в программе "Конспект". Использование автоматизированного метода построения тезауруса терминов предметной области с помощью системы анализа естественно-языковых текстов позволяет значительно сократить затраты времени на составление и редактирование тезауруса. Дальнейшими направлениями исследований является совершенствование алгоритма выделения именных словосочетаний для поиска терминов произвольной длины на основе определенных поверхностных семантических отношений без явного задания всех возможных шаблонов структуры терминов, выделение некоторых типов глагольных словосочетаний и использование их как синонимов к именным.

Благодарности

Работа опубликована при финансовой поддержке проекта **ITHEA XXI** Института информационных теорий и приложений FOI ITHEA Болгария www.ithea.org и Ассоциации создателей и пользователей интеллектуальных систем ADUIS Украина www.aduis.com.ua.

Литература

1. Ноэл М., Спенс К. Microsoft SharePoint 2007 Полное руководство. Издательство: Вильямс, 2008 г. 832 с.
2. Найханова Л.В. Основные аспекты построения онтологий верхнего уровня и предметной области // В сборнике научных статей “Интернет-порталы: содержание и технологии”. Выпуск 3. / Редкол.: А.Н. Тихонов (пред.) и др.; ФГУ ГНИИ ИТТ “Информатика”. – М.: Просвещение, 2005. – С. 452-479.
3. Белоногов Г.Г., Кузнецов Б.А. Языковые средства автоматизированных информационных систем. М.: Наука, 1983.
4. B. Dobrov, N. Loukachevitch, O. Nevzorova. The technology of new domains’ ontologies development // Proceedings of the X-th International Conference “Knowledge-Dialogue-Solution” (KDS’2003).- Varna, Bulgaria.-2003.- pp.283-290.
5. Палагін О.В., Світла С.Ю., Петренко М.Г., Величко В.Ю. Про один підхід до аналізу та розуміння природномовних об’єктів. Комп’ютерні засоби, мережі та системи. -2008, №7. с.128-137.

Информация об авторах

Виталий Величко *– Институт кибернетики имени В.М. Глушкова Национальной академии наук Украины, кандидат технических наук, доцент; проспект Академика Глушкова, 40, Киев-187, 03680, Украина; e-mail: velychko@aduis.com.ua*

Павел Волошин *– ЗАО "Софтлайн", 03142, г. Киев, ул. В.Стуса, 35/37, аналитик компьютерных систем, e-mail: pavelv@unicyb.kiev.ua.*

Светлана Свитла *– Институт кибернетики имени В.М. Глушкова Национальной академии наук Украины, научный сотрудник; проспект Академика Глушкова, 40, Киев-187, 03680, Украина; e-mail:ssve@i.ua*