ТЕЗАУРУС И АВТОМАТИЧЕСКОЕ КОНЦЕПТУАЛЬНОЕ ИНДЕКСИРОВАНИЕ В УНИВЕРСИТЕТСКОЙ ИНФОРМАЦИОННОЙ СИСТЕМЕ “РОССИЯ”

Б.В.Добров, Н.В.Лукашевич

НИВЦ МГУ; АНО Центр информационных исследований
119899, Москва, Воробьевы горы, НИВЦ МГУ, 339
dobroff@mail.cir.ru; *louk@mail.cir.ru*

Введение

Большинство существующих информационно-поисковых систем имеют развитые средства контекстного поиска документов с учетом морфологической информации о словах. Однако в настоящее время очень незначительное число информационных систем предоставляют возможность тематического поиска, например, поиска с использованием тезауруса.

Во многом это связано с тем, что традиционные информационно-поисковые тезаурусы разрабатывались для ручного индексирования человеком-индексатором, а объем потоков информации в настоящее время значительно превосходит возможности индексаторов по их тематической обработке. Как представляется, новым шагом, который мог бы возродить тезаурусный поиск в широком круге информационных систем, является разработка нового типа тезаурусов - тезаурусов, специально разрабатываемых для автоматического индексирования документов.

С 1994 года в АНО Центр Информационных Исследований ведутся работы по разработке Тезауруса для автоматического индексирования текстов общественно-политической тематики. С 1995 года Общественно-политический тезаурус активно и успешно применяется для различных приложений автоматической обработки текстов [1], таких как автоматическое концептуальное индексирование, автоматической рубрицирование с использованием нескольких рубрикаторов, автоматическое аннотирование текстов. Общественно-политический тезаурус - базовый поисковый инструмент в Университетской информационной системе “РОССИЯ” [2] (УИС РОССИЯ – http://www.cir.ru), поддерживаемой НИВЦ МГУ им. М.В.Ломоносова и АНО Центр информационных исследований.

В настоящее время Общественно-политический Тезаурус включает порядка 58 тысяч терминов и наименований, более 25 тысяч понятий, около 100 тысяч отношений между понятиями (около 700 тысяч отношений с учетом иерархии, то есть каждый концепт связан в среднем с 25 другими концептами).

В работе рассматриваются основные отличия Общественно-политического тезауруса для автоматического индексирования от традиционных информационно-поисковых тезаурусов, описываются основные этапы автоматического концептуального индексирования текстов на основе Общественно-политического тезауруса и его функционирование в составе УИС РОССИЯ.

Основные отличия тезауруса для автоматического индексирования от традиционного тезауруса для ручного индексирования

Основной целью разработки традиционных информационно-поисковых тезаурусов [3, 4, 5] является использование их единиц (дескрипторов) для описания основных тем документов в процессе ручного индексирования. По своей сути тезаурус для ручного индексирования является искусственным языком описания, построенным на основе естественного языка. При этом сам процесс индексирования по такому тезаурусу базируется на лингвистических, грамматических знаниях, а также знаниях о предметной области, которые имеются у профессиональных индексаторов текстов. Индексатор сначала должен прочитать текст, понять его и затем изложить содержание текста, пользуясь дескрипторами, указанными в информационно-поисковом тезаурусе. Индексатор должен хорошо понимать всю терминологию, использованную в тексте, - для описания основной темы текста ему понадобится значительно меньшее количество терминов.

При автоматической обработке текстов человека-посредника между текстом и описанием его содержания в виде дескрипторов нет. Есть только автоматический процесс и Тезаурус, который должен содержать и те знания, которые содержатся в традиционных информационно-поисковых тезаурусах, и те знания (насколько это возможно), которые использует индексатор для определения основной темы текста.

Именно поэтому традиционные тезаурусы, разработанные для ручного индексирования, невозможно использовать при автоматическом индексировании [6].

Разработка тезауруса для автоматического индексирования (далее – АИ-тезауруса) характеризуется, прежде всего, необходимостью описания значительно большего количества слов и словосочетаний, встречающихся в текстах данной предметной области. АИ-тезаурус должен включать не только термины, которые представляют важные понятия в текстах данной предметной области, но также охватывать широкий круг более специфических терминов, обнаружение которых в конкретном тексте сделает этот текст релевантным запросу по понятиям более высокого уровня, например, должны быть описаны не только дескриптор *РЫБА* и его основные подразделения, такие как *МОРСКИЕ РЫБЫ*, *АНАДРОМНЫЕ РЫБЫ* и т.п., но и значительное количество конкретных видов рыб с тем, чтобы текст, обсуждающий проблемы вылова минтая, мог бы быть получен при поиске по термину *рыба*.

Синонимические ряды понятий должны быть значительно богаче, чем совокупности вариантов дескриптора в тезаурусе для ручного индексирования, поскольку синонимы должны описывать различные способы выражения данного понятия в тексте для автоматического процесса, а не для человека. Ряды синонимов включают в себя не только существительные и именные группы, а также прилагательные, глаголы, глагольные группы. Расширение терминологической базы АИ-тезауруса ведет к необходимости описания многозначных терминов.

Расширение понятийной базы тезауруса ведет к увеличению и усложнению функций отношений между понятиями тезауруса (концептуальными отношениями): возникает необходимость логического вывода отношений, поскольку описать отношения всех дескрипторов со всеми близкими дескрипторами АИ-тезауруса становится трудоемким занятием и затрудняет проверку таких описаний. В Таблице 1 представлены сравнительные характеристики Общественно-политического тезауруса для автоматического концептуального индексирования и одного из наиболее известных тезаурусов - тезауруса Исследовательской Службы Конгресса США [4], который относится к той же сфере общественных отношений.

Таблица 1. Количественные характеристики состава Общественно-политического тезауруса для автоматического индексирования и Тезауруса Исследовательской службы Конгресса США (LIV)

Характеристика	Общественно-политический Тезаурус	LIV [4]
Число понятий	25 тысяч	6.8 тысяч
Число терминов	58 тысяч	9.8 тысяч
Термины, описанные как многозначные	1.5 тысяч	Нет
Общее количество описанных отношений между понятиями	95 тысяч	15 тысяч
Количество отношений, полученных по логическим свойствам	700 тысяч	Не определено

Общественно-политический тезаурус

Общественно-политический тезаурус представляет собой иерархическую сеть понятий, построенную специально как инструмент для различных приложений автоматической обработки текстов. Тезаурус содержат термины из широкой предметной области общественной жизни, охватывает терминологическую и лексическую информацию, лингвистические и энциклопедические знания, необходимые для анализа содержания нормативных актов Российской Федерации с 1990 г., материалов российских средств массовой информации.

Тезаурус обладает следующими основными особенностями:

- Тезаурус имеет относительно простую систему отношений между понятиями. Основные отношения традиционны: ВЫШЕ -- НИЖЕ, ЦЕЛОЕ -- ЧАСТЬ и АССОЦИАЦИЯ. Кроме того, в отношении АССОЦИАЦИЯ выделено шесть дополнительных отношений с ограниченным наследованием свойств по иерархии. Простая система отношений позволила описать непротиворечивые взаимосвязи между большим числом понятий. Значительные усилия были сделаны для того, чтобы создать многоаспектные классификации понятий, представить разные точки зрения на одно и то же понятие: *ПРИРОДНЫЙ ГАЗ* – это одновременно *ГАЗОБРАЗНОЕ ВЕЩЕСТВО* и *ТОПЛИВО*;
- понятия тезауруса имеют обширные синонимические ряды, включающие много словосочетаний, например, для понятия *ОХРАНА ПРИРОДЫ* - *защита природы*, *природоохранный*, *природозащитный* - всего 24 синонима). Большие ряды синонимов позволяют распознавать понятия Тезауруса в различных контекстах, а также существенно помогают в правильном разрешении многозначных терминов (*печать* либо как *ПЕЧАТЬ НА ДОКУМЕНТ* -- *гербовая печать, круглая печать*; либо как *СМИ* -- *центральная печать*, *периодическая печать*; либо как *ПРОЦЕСС ПЕЧАТИ* -- *офсетная печать*);
- важнейшей особенностью является интеграция тезауруса в процесс автоматической обработки текстов, что позволяет организовать обратную связь, анализируя результаты обработки. Тезаурус постоянно проверяется, подправляется и пополняется по результатам обработки реальных текстов – построенным рубрикам, аннотации, тематическому индексу.

Автоматическое концептуальное индексирование
на основе Общественно-политического тезауруса

Процесс автоматического концептуального индексирования текстов выявляет не 5--10 дескрипторов, выражающих, по мнению человека-индексатора, основную тему документа, а значительно большее количество понятий, обсуждавшихся в документе. Это требует точного определения значимости каждого термина для содержания текста.

В УИС РОССИЯ значимость термина для содержания текста определяется в результате построения “тематического представления текста”, слабо зависящем от величины и типа текстов. В тематическом представлении понятия, обсуждавшиеся в тексте и инициализируемые при обнаружении термина из синонимического ряда понятия, разбиты на совокупности близких по смыслу терминов, так называемые, “тематические узлы”. Тематические узлы текста делятся на “основные узлы”, соответствующие основной теме документа, “локальные узлы”, соответствующие подтемам документа, и “упоминавшиеся”.

Объединение понятий текста в тематические узлы производится на основе тезаурусных связей этих понятий. Например, тема научного исследования может развиваться в тексте посредством следующих понятий: *математика*, *физика*, *прикладное исследование*, *фундаментальное исследование*, *научный работник*.

Классификация тематических узлов на основные, локальные и упоминавшиеся производится на основе фундаментальных свойств связного текста, а именно, его глобальной лексической связности. Действительно, понятия, которые соответствуют основной теме документа, должны проходить через весь текст “красной нитью”. Такую особенность понятий основной темы (основных тематических узлов) можно вычислять алгоритмически [7], и тем самым выделить эти основные тематические узлы из всей совокупности тематических узлов, построенных для текста. Локальные тематические узлы моделируют локальные темы документа, развивающие некоторые из основных тем, то есть связаны с одними основными темами и не связаны с другими.

Совокупность выявленных тематических узлов в тексте образует его тематическое представление. Тематическое представление текста - это иерархическая структура терминов текста, в которой тематически близкие термины собраны вокруг тематических центров в тематические узлы, а среди тематических узлов выделены основные тематические узлы, тематические центры которых отражают основное содержание текста. Тематические узлы связаны между собой отношением *иметь_отношение_к*.

Иерархия тематического представления отражает важность для текста тех или иных терминов. Тематический центр значимее других терминов тематического узла, термины основных тематических узлов более значимы для текста, чем термины других тематических узлов.

Основными этапами формирования тематического представления текста являются:

1. Сопоставление текста с Тезаурусом создает для текста понятийный индекс, в котором указывается, какие понятия Тезауруса и в каком месте текста обнаружены;
2. Нахождение для каждого понятия текста тематически близких (то есть близких по Тезаурусу) понятий и отражение этой информации в так называемой тезаурусной проекции текста;
3. Разрешение многозначности терминов на основе связей понятий в тезаурусной проекции;
4. Построение текстовых связей для каждого понятия текста, то есть фиксация для каждого вхождения каждого понятия трех соседних понятий вправо и трех влево. Выбор таких цифр

величина экспериментальная, однако согласуется и с экспериментами в области исследования кратковременной памяти;

5. Выбор центров тематических узлов. Центрами тематических узлов становятся те понятия, которые отличаются от тематически близких им понятий текста своей частотностью или местоположением в заголовках или начале текста;
6. Выбор среди построенных тематических узлов - основных тематических узлов, то есть тех, которые соответствуют основной теме документа. Выбор производится на основе анализа суммированных текстовых связей тематических узлов.

В результате построения тематического представления текста все термины разделяются на пять базовых классов значимости для текста, каждый из которых имеет свой вес:

- центры основных тематических узлов – 0.95;
- другие понятия основных тематических узлов – 0.85;
- центры локальных тематических узлов – 0.70;
- другие понятия локальных тематических узлов – 0.65;
- упоминавшиеся понятия, не вошедшие в предыдущие классы – 0.20.

Поскольку базовый вес понятия получен в качестве интегрального анализа распределения в тексте совокупностей близких по смыслу терминов, то для получения окончательного веса понятия необходимо учесть относительную частотность этого понятия в тексте. Окончательный вес понятия в тексте $\mu(c, \mathrm{D})$ рассчитывается по следующей формуле:

$$\mu(c, \mathrm{D}) = \lambda \cdot \nu^*(c, \mathrm{D}) + (1-\lambda) \cdot freq(c, \mathrm{D}) \cdot [freq^*(\mathrm{D})]^{-1}$$

где $\nu^*(c, \mathrm{D}) = \max_{Th(c,\ \mathrm{D})} \nu(\mathrm{c}, \mathrm{D})$ – максимум базовых весов понятия c в тематических узлах; оптимальная величина $\lambda = 0.7$; $freq(c, \mathrm{D})$ – частота понятия c в документе D, $freq^*(\mathrm{D}) = \max_{d \in \mathrm{D}} freq(d, \mathrm{D})$ – максимальная частотность среди понятий документа D.

Общественно-политический тезаурус
как инструмент автоматической обработки текстов

С 1995 года Общественно-политический тезаурус используется в таких областях автоматической обработки текстов как автоматическое концептуальное индексирование [7], автоматическая рубрикация текстов [8], автоматическое аннотирование текстов [9]. Все эти применения тезауруса базируются на разработанном авторами тематическом представлении текста.

Тезаурус и технология автоматического построения тематического представления содержания документа позволили развить в рамках УИС РОССИЯ гибкую технологию эффективной автоматической рубрикации текстов. Наши системы автоматической рубрикации работают с такими рубрикаторами как рубрикатор исследовательской службы конгресса США [4], общеправовым тематическим классификатором Центральной избирательной комиссии РФ, классификатором правовых актов РФ [10]. Всего было внедрено шесть различных систем автоматической рубрикации с разными рубрикаторами размером от 35 до 1200 рубрик.

Знания, описанные в Тезаурусе, а также технология построения тематического представления позволили создать систему автоматического аннотирования текстов, основанную на знаниях. В 1998 году с нашей программой автоматического аннотирования англоязычных текстов мы участвовали в соревнованиях

в рамках конференции SUMMAC [11], где эта программа получила лучшие результаты в номинации “Индикативная аннотация наилучшей длины”.

Тезаурус используется как инструмент для автоматического концептуального индексирования и ранжированного информационного поиска в Университетской информационной системе РОССИЯ (www.cir.ru).

Тезаурус как поисковый механизм УИС РОССИЯ

Тематические представления были автоматически построены для текстов различных размеров и жанров. Тематические представления были созданы для более чем 900 Мбайт русских текстов (официальные документы РФ 1990--2000гг., стенограммы пленарных заседаний ГосДумы ФС РФ, сообщения информационных агентств и газетные статьи).

Тезаурус существенно используется в интерфейсе УИС РОССИЯ для следующих задач терминологического поиска:

- уточнения запроса, когда выбор более точного термина позволяет получать только требуемые документы, например, выбирая вместо всех типов *СТРОИТЕЛЬСТВА* именно *ДОРОЖНОЕ СТРОИТЕЛЬСТВО* (*автодорожное строительство, дорожно-строительные работы, строительство дорог, строительно-дорожный* и т.д.);
- автоматического расширения запроса по синонимам (*НАЛОГОВАЯ СИСТЕМА == налоговый режим*), а также по иерархии (*МИГРАЦИЯ НАСЕЛЕНИЯ---- БЕЖЕНЦЫ*, *ВЫНУЖДЕННЫЕ ПЕРЕСЕЛЕНЦЫ* и т.д.).

Тестирование эффективности информационного поиска на основе Тезауруса

Для тестирования эффективности информационного поиска мы выполнили набор запросов в УИС РОССИЯ. Каждый запрос был сформулирован дважды: один раз как запрос на поиск по словам (реализован на основе векторной модели [6], второй раз - как запрос на поиск по понятиям тезауруса с расширением по дереву. В качестве запросов были выбраны рубрики из Классификатора правовых актов [10]. Поиск осуществлялся на 40 тысячной коллекции нормативных актов УИС РОССИЯ.

При выполнении подавляющего количества запросов количество документов, найденных с использованием деревьев Тезауруса значительно превышало количество документов, найденных по словам. Полнота поиска с использованием деревьев тезауруса значительно возросла. Однако, как известно, увеличение полноты поиска часто сопровождается снижением точности поиска, то есть релевантными считается большее количество нерелевантных документов.

Чтобы сопоставить точность поиска по Тезаурусу и по словам на основе векторной модели, мы использовали методику оценки средней точности по трем заданным значениям полноты, описанную в [12]. Точность выполнения запроса вычисляется при следующих трех значениях полноты: 0.2, 0.5, 0.8.

Чтобы оценить эффективность поиска, необходимо сначала определить множество релевантных документов, а затем проверить релевантность значительного количества полученных по запросу документов. Для снижения трудозатрат, необходимых на проведение оценок, мы сохранили формулировку

запроса, но стали сокращать временной интервал до тех пор, пока не получили как релевантные 20-40 документов. Эффективность поиска на таком количестве документов уже достаточно просто проверить.

Приведем результаты наших оценок для одного из запросов. Мы выполнили запрос "Охрана труда" по нормативным документам во временном интервале 01.10.2000 - 01.01.2001 и получили 26 документов при поиске по Тезаурусу и 33 документа при поиске по словам. Просмотрев все полученные документы, мы выяснили, что имеется 28 релевантных документов, причем при поиске по Тезаурусу было найдено 25 релевантных документов, а по словам - 21 релевантный документ.

Точность нужно было вычислить при достижении в списке документов 6-го (6/28=0.2), 14-го (14/28=0.5) и 22-го (22/28=0.8) релевантных документов. При поиске по Тезаурусу шестой релевантный документ был получен седьмым, четырнадцатый - пятнадцатым, двадцать второй -двадцать третьим. Таким образом, средняя точность выполнения запроса по терминам: (0.86+0.93+0.99)/3=0.93.

При поиске по словам шестой релевантный документ был получен десятым, четырнадцатый - двадцать вторым, двадцать второй -не был получен, поскольку было найдено 21 релевантных документов. Средняя точность поиска по словам - (0.60+0.66+0.00)/3=0.42

В настоящее время мы провели тестирование поиска по 14 запросам. Средняя точность поиска, вычисленная для 14 запросов, составляет по терминам - 0.68 (по трем значениям полноты 0.84 - 0.65 - 0.55), по словам - 0.49 (0.77 - 0.67 - 0.02).

Благодарности

Эта работа частично поддержана Российским гуманитарным научным фондом (грант N 00--04--00272а) и Российским Фондом Фундаментальных исследований (грант N 99--06--80107).

Заключение

Описаны особенности и функционирование Общественно-политического тезауруса, разработанного как инструмент автоматической обработки текстов в широкой общественно-политической области.

Опыт разработки и использования Общественно-политического тезауруса позволяет нам утверждать, что и для других тематических коллекций документов могут быть созданы тезаурусы для автоматического концептуального индексирования, которые позволят обеспечить тематический поиск в этих коллекциях.

Литература

1. Loukachevitch N.V.; Salii A.D., Dobrov B.V., Thesaurus for Automatic Indexing: Structure, Developement, Use // Proceedings of International Congress “Terminology and Knowledge Engineering”, 1999, pp. 343--355.
2. Т.Н. Юдина, С.В. Журавлев, “Российский межуниверситетский ресурсный и аналитический центр по гуманитарным исследованиям”, Вестник РФФИ, 1999, N 3 (специальный выпуск), “Наука и информационное общество” (193.233.79.157/pub/vestnik/V3_99/2_8.htm)
3. Шемакин Ю. И. Тезаурус в автоматизированных системах управления и информации. - М: Военное изд-во министерства обороны СССР, 1974. – 192 с.

4. LIV (Legislative Indexing Vocabulary). Congressional Research Service. The Library of Congress. Twenty-first Edition, 1994. 546 p.
5. UNBIS Thesaurus, English Edition, Dag Hammarskjold Library of United Nations, New York, 1976.
6. Salton G., Automatic Text Processing - The Analysis, Transformation and Retrieval of Information by Computer. Addison-Wesley, Reading, MA, 1989.
7. Добров Б.В., Лукашевич Н.В., Использование тематического представления содержания текста для автоматической обработки документов // V Нац. конф. по искусственному интеллекту. - Казань, 1996.
8. Лукашевич Н.В., Автоматическое рубрицирование потоков текстов по общественно-политической тематике // НТИ. Сер.2. - 1996. - N 10. - С.22--30.
9. Лукашевич Н.В., Автоматическое построение аннотаций на основе тематического представления текста // Тр. международного семинара Диалог'97. - Москва, 1997 - С. 188--191.
10. О классификаторе правовых актов, Указ Президента РФ, N 511 от 15 марта 2000г.
11. Tipster SUMMAC Text Summarization Evaluation. Final report. - MITRE Technical report MTR 98000138. - October, 1998. (http://www.itl.nist.gov/iaui/894.02/related_projects/tipster_summac/summac-final-report-part2.ps)
12. Voorhees E., Natural Language Processing and Information Retrieval. // M.T. Pazienza, (Ed.), Information Extraction: Towards Scalable, Adaptable Systems, Germany: Springer, 1999, pp.32-48.

THESAURUS AND AUTOMATIC CONCEPTUAL INDEXING IN UNIVERSITY INFORMATION SYSTEM "RUSSIA"

Boris V. Dobrov, Natalia V. Loukachevitch
dobroff@mail.cir.ru; louk@mail.cir.ru
Research Computer Center of Moscow State University; NCO Center for Information Research

The paper describes the structure of the Sociopolitical thesaurus, which was specially created as a tool for automatic conceptual indexing.
We compare main features of the Sociopolitical thesaurus and traditional information-retrieval thesauri. Evaluation of document search based on the Sociopolitical thesaurus shows considerable increase of retrieval effectiveness.