Автоматическое построение терминологической базы знаний♣

© О.Г. Чанышев

ОФИМ СО РАН
fedorov22@yandex.ru

Аннотация

В докладе представлены метод автоматического построения и структура терминологической базы знаний предметной области, заданной множеством естественно-языковых текстов. Автоматически строится предметный указатель, путем кластеризации терминоподобных словосочетаний по принципу вхождения в них главного («кардинального») слова. Вводится мера ассоциативной близости терминоподобных словосочетаний. Кратко описан процесс выделения терминоподобных словосочетаний. Основным фактором, определяющим их адекватность предметным областям, является доминантность. Приведен результат оценки адекватности.

1 Введение

Несомненный интерес представляют работы по автоматизации построения баз знаний [2, 4, 5] путем экстракции знаний из естественно-языковых текстов, представляющих некоторую предметную область. Практически полезной и не требующей детального семантико-синтаксического анализа предложений для ее построения может быть терминологическая база знаний.

Настоящая работа выполнена в рамках комплекса исследований, посвященных автоматическому анализу естественно-языковых текстов и автоматизации построения семантических сетей отдельных текстов и предметных областей. В ходе исследований были разработаны: Ассоциативная модель реального текста [6, 7], позволившая, в частности, выделять наиболее тематически значимые слова (доминанты), оригинальные алгоритмы распознавания различных фрагментов текстов (рубрик и их наименований, списков литературы, ФИО, границ предложений) [9], простой алгоритм классификации текстов на основе анализа пересечений множеств доминант [5].

Естественным продолжением стала разработка метода автоматического выделения терминоподобных словосочетаний (ТС) [9], для практической реализации которого использовался весь наработанный ранее инструментарий. Главное дополнение к известным из литературы условиям, которым должны отвечать ТС (контактность, устойчивость, объектность (обязательное наличие существительного), смысловая завершенность и ограниченная распространенность), – условие доминантности: каждое словосочетание должно включать слово, принадлежащее множеству доминант как минимум в одном из текстов анализируемого множества. Это условие обеспечивает адекватность выделенных словосочетаний предметной области и повышает эффективность фильтрации.

В докладе акцент смещен в сторону организации информации о терминоподобных словосочетаниях в базе знаний интеллектуального прототипа терминологической ИПС (в дальнейшем – просто ИПС). Структура фактов определяется требованиями к ИПС.

2 Требования к терминологической ИПС

1. Прежде всего, показалась привлекательной идея «педагогического» характера – дать возможность пользователю получить представление о содержании ТС в пределах множества текстов предметной области путем определения и вывода всех предложений вхождения элементов его грамматической парадигмы.

2. Эту возможность следовало дополнить быстрым доступом к энциклопедическим и толковым словарям, в частности, для построения конспектов по «локальным темам», определяемым терминоподобным сочетанием.

3. Представляется, что наиболее удобным для пользователя вариантом выбора ТС является выбор по предметному указателю, в котором ТС кластеризованы в группы по принципу общего главного или «кардинального» слова.

4. Очевидно, следовало добавить сервисы, создание которых не представляет сложностей:

предоставление списка текстов вхождения выбранного ТС для доступа к ним;

Труды 10-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2008, Дубна, Россия, 2008.

поливариантный выбор ТС – по норме, по элементам грамматической парадигмы, по заданной подстроке.

5. Чтобы удовлетворить первому требованию, необходимо автоматически определять вхождения ТС в тексты и предложения текстов, тем самым, создавая дерево вхождений, в котором узлы – тексты, предложения и ТС связаны отношениями включения. Это дерево можно сделать терминологической семантической сетью, если ввести два вида отношений между парами ТС – отношение тематической важности и отношение ассоциативной близости. В таком случае представляется возможность для интерпретации запросов пользователя, сформулированных на естественном языке как запросов на поиск путей между ТС в терминологической семантической сети.

Конечно, на данном этапе исследований интерпретация будет основываться на анализе ключевых слов, в качестве которых выступают ТС. «Анализ ключевых слов есть метод анализа предложений на предмет наличия ключевых слов, которые может распознавать компьютер. Ключевые слова становятся значениями объектов предикатов. При этом не так важна грамматическая структура предложения, так как программа не анализирует связей между словами. Компьютер реагирует одинаково на различные варианты входного текста; роль играет лишь наличие ключевых слов» [3].

Результат может быть представлен пользователю в виде множества предложений вхождения терминов, выделенных из запроса и промежуточных терминов. Учитывая, что каждому терминоподобному сочетанию приписывается тематический вес, появляется возможность сравнить различные пути по двум параметрам: по значению ассоциативной близости и их тематической важности.

Таким образом, в список первоочередных задач входили две основных: задача определения кардинального слова и задача выбора меры ассоциативной близости ТС.

3 Необходимые определения

3.1 Ассоциативная мощность

Несмотря на то, что Ассоциативная модель была неоднократно изложена в ранних публикациях, для лучшего восприятия настоящего текста приведем основные положения, касающиеся понятия «ассоциативная мощность» – основы меры тематической важности ТС.

За установление ассоциативных связей между предложениями текста отвечает множество независимых лексем связи (НЛС), обладающее тем свойством, что для каждой произвольной пары из этого множества существуют минимум два предложения, в которые они входят по отдельности (элементы НЛС не принадлежат стоп-словарю и их частота встречаемости в различных предложениях больше 1).

Мера тематической важности («тематический вес») лексемы из НЛС (*Ш* – **ассоциативная мощность)** равна числу других лексем из НЛС, встречающейся с данной в предложениях текста, с поправкой, учитывающей “личный” вклад лексемы в установлении связей – лексемы первого предложения вхождения не учитываются, поскольку в нем присутствуют только «родители» и «братья» данной.

Из НЛС отбираются **доминантные лексемы** (доминанты) по критерию $Ш>0.5R$, где R – ранг текста (максимальный номер группы лексем с равными *Ш* в частично упорядоченной по убыванию *Ш* последовательности НЛС).

3.2 Нормы и парадигмы ТС

Норма (лемма) слова. Норма для каждого слова словосочетания определяется по строке морфологической информации, выдаваемой морфопарсером mystem.

Норма ТС получается заменой слов на леммы.

Парадигма ТС (слова) – множество грамматических вариантов ТС (слова) в пределах заданного множества текстов. Размер парадигмы – число вариантов.

3.3 Тематические веса

Для работы с множествами текстов в**ес доминанты** в фиксированном тексте полагается равным ее обратному рангу в убывающей по значению ассоциативной мощности последовательности доминант.

Вес нормы доминанты во множестве файлов равен сумме весов ее доминантных грамматических форм.

Вес словосочетания равен сумме весов входящих доминант.

Вес нормы сочетания равен сумме весов элементов его парадигмы.

3.4 Кардинальные слова и предметный указатель

Кардинальные слова

Кардинальным словом нормы словосочетания назовем норму доминанты с наибольшим весом. По нашей (вероятно, наивной) оценке кардинальные слова – это однословные термины, не обязательно принадлежащие рассматриваемой предметной области. В Таблице 1 для шести предметных областей приводятся данные авторской оценки «термин – не термин».

Метод оценки: не вдаваясь в анализ принадлежности слова к специализированной или общеупотребительной лексике, в качестве терминов рассматривались существительные и аббревиатуры. Редко – прилагательные, явно выступающие в роли существительных в конкретном тексте. Например, *больная, испытуемый* в рамках предметной области «Психология».

Ниже приводится полный список кардинальных слов предметной области «Компьютерная лингвистика». Кардинальные слова упорядочены по убыванию весов. Не термины набраны жирным курсивом.

слово, документ, текст, знание, данный, модель, цепочка, правило, система, обработка, отношение, термин, проектирование, информация, рубрикация запрос, статья, анализ, группа, словарь, выборка предложение, ***строить****, ооо, wordnet, словосочетание,* ***позволять****, ошибка, вид,* ***объектовый****, построение, поиск,* ***контекстовый****, язык, объект, понятие, качество, сеть, работа, вывод, связь,* ***значимый****,* ***называть****, грамматика, время, связанность, смысл,* ***семантический****, medsearch, сапр, управление, процесс, длина, number, технология, число, база, значимость,* ***иметь****, рубрика,* ***ассоциативный****, встречаемость, класс, рассуждение, савс,* ***лекарственный****, tin, представление,* ***логический****, эксперт, логика,* ***конкретный****, рубрицирование, произведение, множество, проверка,* ***являть****, структура, тип,* ***макаревич****, функция, описание, решение, образ, создание, разработка, единица, день,* ***максимальный****,* ***языковый****, эталон, бд, text, метод, память,* ***and****, шаблон, лексема,* ***электронный****,* ***the****, средство,* ***именной****,* ***второй****, предприятие,* ***нетрадиционный****, диалог, продукция, сотрудник,* ***начальный****, интернет,* ***лингвистический****, значение, пример, человек,* ***следующий****, алгоритм,* ***научно-технический****, oracle,* ***комбинаторный, реальный, ап****, окрестность,* ***описывать****, моделирование,* ***один****, запись, изделие, пользователь,* ***стоить****,* ***курейчик****,* ***московский****, обеспечение, сервер,* ***деловой****, доступ, производительность, компания, задание.*

Таблица 1. Результаты оценки термин – не термин для кардинальных слов

Предметн. обл.	Число кард. слов	Число не терминов	Отношение
Иск. интел.	172	33	0.19
Сетевые операц. системы	162	32	0.20
СУБД	168	33	0.20
Психология	234	62	0.26
Философия	329	105	0.32

Построение предметного указателя

Кардинальные слова упорядочиваются по убыванию их тематических весов. Все нормы ТС подразделяются на группы по признаку вхождения в них кардинальных слов. Нормы ТС в группах упорядочиваются по возрастанию их длины (длина - число слов). В группах организуются подгруппы по признаку включения более коротких ТС.

Пример (для лучшей читабельности ТС приведены вручную к номинативной форме):

система
система ии
современная система ии
построение системы ии
история развития систем ии
система искусственного интеллекта
совершенствование системы искусственного интеллекта
современная система искусственного интеллекта
система понимания естественного языка

В результате группирования часть кардинальных слов может остаться без своих включающих словосочетаний. В таком случае для них организуются ссылки на соответствующие группы. Пример:

понимание->система->система понимания естественного языка

По сути – это алгоритмически простой вариант иерархического группирования терминов. Более сложные варианты рассмотрены в работе [1].

3.5 Контекстная мера ассоциативной близости

Для групп ТС, идентифицированных кардинальными словами Ki,Kj – это функция от трех параметров: N - числа общих текстов (содержащих хотя бы по одному элементу парадигмы из состава терминов различных групп), Lmin – минимального расстояния между предложениями, включающими элементы парадигм различных групп и L – среднего расстояния между предложениями во всем множестве общих текстов.

Например:

A(Ki,Kj)=aN/(1+LЧLmin), a – нормировочный коэффициент.

Очевидно, эту же меру можно использовать и для определения близости пары норм терминоподобных словосочетаний. В таком случае L и Lmin – это среднее и минимальное расстояния между предложениями, включающими элементы парадигм норм ТС.

Тематическая важность группы ТС определяется как сумма весов норм ТС, входящих в группу.

4 Формирование терминологической базы знаний

4.1 О выборе языков программирования

Для реализации алгоритма формирования базы знаний (автоматического анализа текстов, определения словосочетаний и итогового формирования фактов в синтаксисе Пролога) выбран свободно распространяемый язык программирования Unicon (расширенный Icon), а прототип ИПС реализован на PDC Prolog v.5.2. Personal Edition.

Выбор основан на следующих соображениях: реализация процесса автоматического анализа текста с целью выделения ТС требует большого объема исходного кода. Язык Unicon обладает очень простым синтаксисом и чрезвычайно удобными и функционально мощными встроенными типами данных (таблицы, множества, списки, записи) и полным набором стандартных математических операций и встроенных функций. Компилятор Unicon строит программы, по скорости работы не усту-

пающими программам, реализованными на Си. Unicon мог бы быть идеальным языком быстрого прототипирования, если бы не проблемы с кириллицей в интерактивном режиме работы.

Основываясь на собственном многолетнем опыте работы с Prolog'ом, можем утверждать, что система программирования PDC Prolog v.5.2., удовлетворяя определению Объектно-ориентированных баз данных (СУБД+язык программирования) очень удобна для экспериментов с ИПС в силу логической простоты и высокой технологичности. В частности, организация всякого рода многоуровневых меню, управления процессами не представляет никаких сложностей из-за наличия множества специализированных встроенных предикатов. Основные усилия должны быть сосредоточены на организации баз данных. Данные могут быть представлены в виде простой и удобной системы внутренних баз данных, которые содержат факты Пролога. Базы могут модифицироваться, сохраняться в файлы и загружаться в оперативную память из файлов. Для коммерческих систем с очень большим количеством информации следует использовать систему внешних (бинарных) баз данных.

Таким образом, комбинация этих двух языков программирования полностью удовлетворяет потребностям исследователя при разработке прототипов интеллектуальных систем («экспериментальных стендов»), избавляя от необходимости обращаться к профессиональным программистам.

4.2 Основные этапы выделения ТС

На вход программной системы поступает список полных имен тестовых файлов, представляющих фиксированную предметную область (или некоторый объемный текст).

1. В каждом текстовом файле программа выделяет и классифицирует лексемы и строки. Затем, на основании полученной признаковой информации, распознает границы предложений.

2. Выделяются НЛС и доминанты.

3. Параллельно (последовательность не важна) определяются все возможные контактные устойчивые словосочетания. Если частота контактного словосочетания больше единицы, то в справочниках соответствующих предложений вхождения эта пара лексем заменяется на их сочетание и новая «лексема» добавляется в словарь текста. Процесс повторяется для лексемы-сочетания.

4. Используя морфоанализатор mystem, для каждого слова сочетания получаем морфологическую информацию.

5. Словосочетания отдельных файлов объединяются. Определяются веса словосочетаний, затем они группируются в парадигмы. Каждому элементу парадигмы сопоставляется список включений, содержащий наименования текстов, полные имена файлов и списки предложений вхождения для каждого файла.

6. **Удаляются словосочетания, не содержащие доминант**.

7. Оставшиеся словосочетания пропускаются через лексико-морфологический фильтр.

8. Пересчитываются веса словосочетаний и их норм.

9. Формируются файлы фактов предметной области в синтаксисе Prolog'а.

4.3 Коэффициенты фильтрации

В настоящее время база знаний содержит описания следующих предметных областей (и одной монографии, представленной десятью файлами):

«СУБД», «Искусственный интеллект», «Философия», «Психология», «Сетевые операционные системы» (монография).

Количества текстов предметных областей (в порядке перечисления): 13,18,54,98,10,15.

Коэффициенты фильтрации (значение отношения числа всех отвергнутых словосочетаний к общему числу выделенных):

Психология – 0.39
Искусственный интеллект – 0.37,
Сетевые операционные системы – 0.34,
СУБД – 0.30,
Философия – 0.28.

Процент малоинформативных словосочетаний в итоговых множествах (по визуальной субъективной оценке) не превышает 2%.

Отметим, что в приведенном макроалгоритме принципиальной новизной обладает только п.6, выполнение которого обеспечивает адекватность выделенных словосочетаний предметным областям.

4.4 Оценка адекватности ТС предметным областям

Адекватность выделенных ТС предметным областям оценивалась путем классификации множеств норм словосочетаний, выделенных нами (контрольные множества), по предметным областям, представленным наименованиями статей словарей (эталонные множества словосочетаний) [9]. Принадлежность контрольных множеств к ПО определялась по максимальному суммарному весу норм словосочетаний, принадлежащих пересечению с эталонными множествами.

В качестве эталонных множеств словосочетаний были выбраны:

а) нормированные наименования статей "Новейшего философского словаря под редакцией Грицанова А.А." (http://linguists.narod.ru/downloads5.html#spec, 1390 наименований, "Философия-эталон");

б) нормированные термины "Психологического словаря" (http://psi.webzone.ru/, 2172 наименования, "Психология-эталон").

в) в качестве эталонного множества сочетаний, представляющих информатику, были выбраны нормированные наименования статей "Словаря компьютерной лексики" (http://slovar.boom.ru/Head.html, 1213 наименований, "КомпЛекс-эталон").

В контрольные множества словосочетаний были включены нормы словосочетаний “СУБД”, “СетОпСист”, “Иск. Инт.” “Философия”, “Психология”. А также для контроля качества подборок мы проанализировали “Психологические теории и концепции личности. Краткий справочник.” (http://www.gumer.info/bibliotek_Buks/Psihol/Psi_Teo/index.php)
и нормированные двухсловные словосочетания включили в контрольную подборку (“ПсихТеор”).

Результат представлен в Таблице 2. Первая строка в каждой ячейке – результат до фильтрации, вторая – после фильтрации. Первое число в строке – суммарный вес словосочетаний контрольной подборки, принадлежащих множеству пересечения с эталонной подборкой, второе – размер пересечения.

Таблица 2. Результат классификации множеств словосочетаний.

Предм. обл. Эталон	КомпЛекс эталон	Философ. Эталон	Психология Эталон
СУБД	359.05 46 359.05 39		0.00 1
СетОпСист	82.05 69 82.05 56		1.79 2 1.79 1
Иск.Инт.	49.61 13 49.61 10		5.73 5 5.73 4
Философия	0.18 1 0.18 1	184.24 33 184.09 28	7.65 7 7.16 6
Психология	2.53 12 2.27 10	2.47 3 2.47 2	47.83 53 47.83 44
ПсихТеор		2.35 4 2.35 3	28.57 13 28.57 13

4.5 Некоторые характеристики программного обеспечения

Максимальный размер исполняемого модуля – 796 Кб.
Вычислительная система:
Intel(R) Core(TM)2 Duo CPU E6750 @ 2.66GHz
Память - 2 Гб.

Максимальное время – 30 сек для подборки текстов из предметной области «Философия», 54 текста общим объемом – 2 Мб.

5 База знаний

5.1 Структура

Структурно терминологическая база знаний организована в виде семантической сети, основу которой составляет дерево вхождений ТС в тексты и предложения, а группы ТС, идентифицированные кардинальными словами, связаны отношениями ассоциативной близости и тематической важности.

5.2 Файлы

Базу знаний системы составляют текстовые файлы, содержащие тексты предметных областей, тексты энциклопедических и толковых словарей и файлы фактов (метаинформация).

В систему включены следующие словари:

«Словарь компьютерной лексики», «Современный энциклопедический словарь», «Современный философский словарь», «Все монархи мира. Др.Греция, Рим», «Все монархи мира. Западная Европа», «Биографический справочник», «Толковый словарь Ефремовой».

5.3 Файлы, содержащие тексты

Структура текстов в базе знаний отличается от структуры исходных текстов наличием номеров предложений (в угловых скобках, «тэгах») перед каждым предложением. Они добавляются автоматически в процессе распознавания границ предложений.

5.5 Файлы фактов

Множество файлов фактов составляют:

файл состава базы знаний, файл словарного состава, файлы – описатели вхождений ТС, файлы предметных указателей, файлы весов норм ТС.

Для сокращения описания доменного состава фактов введем следующие обозначения:

ПО – предметная область

ННТС – номер нормы терминоподобного словосочетания,

НТС – норма терминоподобного словосочетания,

ЭП – элемент парадигмы,

НТ- наименование текста,

ПИТФ – полное имя файла, содержащего текст,

СПВ – список номеров предложений вхождения,

НСл – наименование словаря.

Файл состава базы знаний содержит множество однотипных фактов

имена_баз(<имя_ПО>,<имя_файла_ПО>),

сопоставляющих наименованию предметной области OS-имя файла, содержащего факты, описывающие связи между нормами словосочетаний, множествами парадигм и множествами включающих предложений различных текстов.

Файл словарного состава содержит множество однотипных фактов, сопоставляющих наименованию словаря имя содержащего файла:

имя_словаря(<НСл>,<ПИТФ>).

Файлы – описатели вхождений ТС содержат факты следующих типов:

норма(<ННТС>,<НТС>)
елемент_парадигмы(<ННТС>,<ЭП>)
файл(<ННТС>,<НТ>,<ПИТФ>,<СПВ>)
Пример:
норма(12,"объект база данный")
елемент_парадигмы(12,"объект базы данных")
елемент_парадигмы(12,"объектом базы данных")
елемент_парадигмы(12,"объекта базы данных")
елемент_парадигмы(12,"объектам базы данных")
файл(12,"С.Д. Кузнецов. Введение в СУБД. Часть7.", "D:\\TEXTS\\SUBD\\bd_kuz7.ntf",[76,99])

Файлы предметных указателей содержат факты следующих типов

группа(<КС><НТС>)

ссылка_на_группу(<КС1><КС2><НТС>),

где *КС* – кардинальное слово.

Пример:

группа("искусственный","искусственный жизнь")

группа("искусственный","база искусственный жизнь")

группа("искусственный","изучение искусственный жизнь")

ссылка_на_группу("жизнь","искусственный","искусственный жизнь")

ссылка_на_группу("жизнь","искусственный","база искусственный жизнь")

ссылка_на_группу("жизнь","искусственный","изучение искусственный жизнь")

Файлы весов норм ТС содержат факты, сопоставляющие нормам ТС их веса:

вес_нормы(<НТС>,<вес_НТС>)

6 Терминологическая ИПС

Файлы фактов загружаются во внутренние базы данных прототипа терминологической информационно-поисковой системы (ИПС).

В настоящее время ИПС позволяет:

- выбирать предметные области и загружать ассоциированные с ней факты;
- вводить или выбирать интересующие пользователя ТС из групп или из списков ТС, отсортированных по алфавиту, либо по убыванию их весов;
- предоставлять пользователю все предложения вхождения элементов парадигмы ТС или группы ТС из всего множества текстов, представляющих выбранную предметную область;
- производить поиск непосредственно в текстах;
- искать необходимую информацию в словарях;
- строить отношения ассоциативной близости и тематической важности между группами идентифицированными кардинальными словами.

При запуске программы прежде всего загружаются файл состава базы знаний и файл словарного состава. Для дальнейшей работы пользователь должен выбрать предметную область. После этого загружаются факты, связанные с данной предметной областью. Имена файлов фактов образуются по правилу

< имя_файла_ПО _без_расширения>.<расширение>

Конструируются факты, задающие отношения ассоциативной близости и тематической важности между группами, идентифицированными кардинальными словами:

а_близость(<КС><значение_а_близости>),

т_важность(<КС><значение_т_важности>)

6.1 Пример использования фактов для вывода всех предложений вхождения элементов парадигмы ТС

1. Задается норма ТС (*НТС*).
2. Используя факт *норма(<ННТС>,<НТС>),* определяется значение *<ННТС>*.
3. Отбираются факты

файл(<ННТС>,<НТ>,<ПИТФ>,<СПВ>)

с заданным значением *<ННТС>*.

4. Из каждого файла с OS-именем *<ПИТФ>* выбираются предложения с номерами, содержащимися в списке *<СПВ>*.
5. Предложения с предшествующими наименованиями текстов (*<НТ>*) объединяются в «строку» и выводятся на экран.

6.2 Использование других фактов

Факт *елемент_парадигмы(<ННТС>,<ЭП>)* используется для поиска значения *<ННТС>* по заданному элементу парадигмы или при поиске *<ННТС>* по заданной подстроке.

Факты

группа(<КС><НТС>),

ссылка_на_группу(<КС1><КС2><НТС>)

используются для выбора *НТС* из групп после предварительного выбора кардинального слова, а также для построения отношений близости и тематической важности между группами, идентифицированными кардинальными словами.

6.3 Примеры вывода предложений вхождения.

Предметная область – СУБД, норма термина – «тип данный».

ТЕКСТ >> С. Д. КУЗНЕЦОВ. ВВЕДЕНИЕ В СУБД. ЧАСТЬ 9.

*<119>Недаром в стандарте SQL появились специальные **типы данных** date и time.*

*<155> Соответствующий базис обеспечивают как предыдущие работы в области БД, так и давно развивающиеся направления языков программирования с абстрактными **типами данных** и объектно-ориентированных языков программирования.*

*<192>Как видно, при таком наборе базовых понятий, если не принимать во внимание возможности наследования классов и соответствующие проблемы, объектно-ориентированный подход очень близок к подходу языков программирования с абстрактными (или произвольными) **типами данных**.*

*<440>Такая информация в случае реляционной базы данных сохраняется в системных отношениях-каталогах и содержит главным образом имена базовых отношений и имена и **типы данных** их атрибутов.*

ТЕКСТ >> С. Д. КУЗНЕЦОВ. ВВЕДЕНИЕ В СУБД: ЧАСТЬ 3

*<22>Основными понятиями реляционных баз данных являются: **тип данных**, домен, атрибут, кортеж, первичный ключ и отношение.*

*<26>4.1.1 **Тип данных**.*

*<27>Понятие **тип данных** в реляционной модели данных полностью адекватно понятию **типа данных** в языках программирования.*

Предметная область – «Философия», норма термина – «право человек».

ТЕКСТ >> ВВЕДЕНИЕ В ФИЛОСОФИЮ. ДУХОВНАЯ ЖИЗНЬ ОБЩЕСТВА.

*<216>В праве сформулированы **права человека**, например, право на сохранение и продолжение жизни, в то же время в праве отражены нравственные обязательства индивида.*

*<230>Во Всеобщей декларации **прав человека** записано: "Каждый человек имеет право на жизнь, на свободу и на личную неприкосновенность".*

*<235>Если же естественные **права человека** ущемлены, то он чувствует себя неуверенно, теряет способность к творчеству, тем самым сдерживая собственное развитие и развитие общества.*

*<237>Незыблемость **прав человека**, его чести и достоинства, его интересов, их охрана и гарантированность являются принципами правового государства.*

Предметная область – «Психология», норма термина – «активация асфс».

ТЕКСТ >> И.СМИРНОВ, Е.БЕЗНОСЮК, А.ЖУРАВЛЁВ. ПСИХОТЕХНОЛОГИИ.

*<2535>Л.В.Бережкова и В.М.Смирнов считают, что АСФС, как феномен долгосрочной памяти, характеризуется специфичным для нее паттерном сверхмедленных колебаний потенциала, который проявляется при **активации АСФС** и сопровождается эффектами уменьшения проявлений основного заболевания и повышением психической активности.*

*<2541>В этой работе указано, что АСФС-II можно формировать двумя путями: простые матрицы, когда на фоне однократного введения этимизола предъявляют сенсорную стимуляцию и последующие сеансы **активации АСФС**-II проводят только с помощью сенсорной стимуляции разной модальности, но равной частоты; и сложные матрицы, когда в структуру АСФС-II включают фармакологические препараты, например, для невротиков используют транквилизаторы, а для больных с фантомно-болевыми синдромами - анальгетики.*

*<2543>У здоровых напряженно работающих лиц через 4-7 сеансов **активации АСФС**-II объем непосредственной памяти возрос на 42%, оперативной - на 77%, индекс кратковременной памяти - на 92% (двойной тест), устойчивость внимания повысилась на 50%, пропускная способность зрительного анализатора возросла до 42% (корректурный тест), оптимизировалось эмоциональное состояние, улучшилась способность к обучению, повысилась психическая активность.*

*<2551>После однократной **активации АСФС**-1 лечебный эффект превышает два года (Бородкин Ю.С., Зайцев Ю.В., 1985).*

6.4 О применимости меры ассоциативной близости

В настоящее время проводятся эксперименты по интерпретации запросов пользователя на семантической терминологической сети. Пока в качестве примера, иллюстрирующего применимость введенной меры ассоциативной близости (А) приведем по 10 ассоциаций с кардинальным словом «память», расположенных в порядке убывания ненормированного значения А, из трех предметных областей.

«СУБД»:

память-журнал 5
память-функция 5
память-файл 4
память-управление 2.86
память-число 1.54
память-страница 0.41
память-объект 0.21
память-организация 0.21
память-значение 0.21
память-кортеж 0.18

«Сетевые операционные системы»:

память-адрес 2.22
память-страница 1.58
память-сетевой 1.46
память-использование 0.68
память-область 0.57
память-пространство 0.50
память-управление 0.35
память-таблица 0.23
память-сервер 0.21
память-сообщение 0.21

«Психология»:

память-асфс 0.59
память-психика 0.16
память-семантический 0.07
память-уровень 0.05
память-информация 0.04
память-мозг 0.04
память-состояние 0.03
память-расстройство 0.03
память-исследование 0.02
память-реакция 0.02

6.5 О временных характеристиках

Из всех операций в данной реализации ИПС самой «времяемкой» оказалась загрузка словаря объемом 13.6 Мб – 3 сек.

Максимальная продолжительность построения отношений близости для всех групп – около 1 сек.

7 Выводы

Построен прототип интеллектуальной терминологической ИПС. Структура его автоматически сгенерированной базы знаний полностью отвечает сформулированным требованиям.

Адекватность выделенных терминоподобных словосочетаний предметным областям обеспечива-

ется, в основном, исключением словосочетаний, не содержащих доминантных лексем.

Реализован относительно простой и эффективный метод автоматического построения предметного указателя на основе выделения «кардинальных слов» словосочетаний.

Возможность представления пользователю множества предложений вхождения выбранного терминологического сочетания является хорошей основой для построения конспектов по «локальным темам» и может использоваться в автоматизированных обучающих системах.

Результаты первых экспериментов позволяют говорить о приемлемости введенной меры ассоциативной близости для интерпретации запросов пользователя как поиска путей между терминологическими словосочетаниями в семантической терминологической сети.

Литература

[1] Виноградова Н.В., Митрофанова О.А., Паничева П.В. Автоматическая классификация терминов в русскоязычном корпусе текстов по корпусной лингвистике // Труды 9-й всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2007, Переславль-Залесский, Россия, 2007.

[2] Добров Б.В., Лукашевич Н.В., Сыромятников С.В. Формирование базы терминологических словосочетаний по текстам предметной области // Труды 5-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2003, Санкт-Перербург, Россия, 2003.

[3] Ин Ц., Соломон Д. Использование Турбо-Пролога : пер. с англ. – М. : Мир, 1993.

[4] Загоруйко Н.Г., Налетов А.М., Гребенкин И.М. На пути к автоматическому построению онтологии (http://www.dialog-21.ru/Archive/2003/Zagorujko.htm)

[5] Чанышев О.Г. О возможности построения онтологий на основе доминантных лексем: результаты автоклассификации текстов // Вестник Омского государственного университета. – 2004. – Вып. 3. – С.45–47.

[6] Чанышев О.Г. Диссертация «Ассоциативная модель реального текста и ее применение для автогенерации баз знаний о связях» на соискание степени кандидата технических наук по специальности 05.13.16. – применение вычислительной техники, математического моделирования и математических методов в научных исследованиях // Омск, ОФИМ СО РАН, 1998 – 120 с.

[7] Чанышев О.Г. Ассоциативная модель реального текста и ее применение в процессах автоиндексирования // Труды Седьмой национальной конференции по искусственному интеллекту с международным участием КИИ'2000. – М. : Физ.-мат. лит., 2000. – С. 430–438.

[8] Чанышев О.Г. Автоматическое извлечение кандидатов в термины предметной области из представляющих ее текстов // Информационные технологии. – 2008, №2. – С. 2–7.

[9] О.Г. Чанышев. О распознавании фрагментов естественно-языкового текста // Вестник Омского государственного университета. – 2002. – Вып. 4. – С. 14–16.

Automatic construction of terminological knowledge base

O.G. Chanyshev

In the report the method of automatic construction and structure of the terminological knowledge base of the subject domain set by set of naturally-language texts are stated. The index, by clusterization term-like word-combinations by a principle of entering in them of the main thing ("cardinal") word automatically construction. The measure of associative affinity term-like word-combinations is entered. Process of extraction term-like word-combinations is briefly described. A major factor defining their adequacy to subject domains is presence most thematically important words - dominants. They also are defined automatically. The result of an estimation of adequacy is given. Also the result of definition of associative affinity of a cardinal word "память"(memory) for three subject domains is given.

♣ Работы выполнена по проекту № 1.4.2. ОМН РАН за 2007 г.