Маркелов Константин Сергеевич, Нейман Андрей Борисович

МЕТОДЫ АВТОМАТИЗИРОВАННОГО СИНТЕЗА ТЕЗАУРУСОВ ПРЕДМЕТНЫХ ОБЛАСТЕЙ НА ОСНОВЕ АНАЛИЗА ТЕКСТОВ

Статья раскрывает понятие тезаурусов предметных областей, а также в ней рассматриваются методы анализа текстов, на основе которых автоматизированно строится словарь - тезаурус любой предметной области. В работе представлена структура алгоритма автоматизированного построения тезауруса, с помощью которой возможна реализация программно-алгоритмического обеспечения.

Адрес статьи: www.gramota.net/materials/1/2012/7/24.html

Источник

Альманах современной науки и образования

Тамбов: Грамота, 2012. № 7 (62). С. 81-86. ISSN 1993-5552.

Адрес журнала: www.gramota.net/editions/1.html

Содержание данного номера журнала: www.gramota.net/materials/1/2012/7/

УДК 681.3
Технические науки

Статья раскрывает понятие тезаурусов предметных областей, а также в ней рассматриваются методы анализа текстов, на основе которых автоматизированно строится словарь - тезаурус любой предметной области. В работе представлена структура алгоритма автоматизированного построения тезауруса, с помощью которой возможна реализация программно-алгоритмического обеспечения.

Ключевые слова и фразы: тезаурус; предметная область; искусственный интеллект; методы анализа текстов; статистический анализ; лингвистический анализ.

Константин Сергеевич Маркелов
Андрей Борисович Нейман
Московский государственный технический университет радиотехники, электроники и автоматики (МГТУ МИРЭА)
kosmar89@mail.ru; deadloko@gmail.com

МЕТОДЫ АВТОМАТИЗИРОВАННОГО СИНТЕЗА ТЕЗАУРУСОВ ПРЕДМЕТНЫХ ОБЛАСТЕЙ НА ОСНОВЕ АНАЛИЗА ТЕКСТОВ©

Введение

Состояние вопроса об автоматизированном формировании тезаурусов предметных областей трудно оценить, так как литературы по этой проблеме достаточно мало. В то же время проблеме уделялось большое внимание, однако каких-то продуктов, позволяющих сформировать тезаурус автоматически, практически нет. Существуют только конкретные пользовательские наработки, позволяющие решать только определенные классы задач или работать только с определенной тематикой предметной области (ПО).

Актуальность темы определяется тем, что тезаурус самому создать достаточно сложно, а созданный автоматически тезаурус поможет при изучении любой предметной области. Такой тезаурус не будет содержать ошибок и не будет опускать какие-то определения и понятия, так как при формировании такого тезауруса используются математические и статистические методы. Правильно сформированный тезаурус будет являться базисом для изучения любой предметной области.

Под ***тезаурусом*** понимается сложный компонент словарного типа, в котором все значения словаря связаны между собой семантическими отношениями, отражающими основные соотношения понятий в описываемой предметной области знаний.

В состав тезауруса входят лексемы, относящиеся к четырем частям речи: прилагательному, существительному, глаголу и наречию. Описания, соответствующие каждой части речи, имеют различную структуру.

Все отношения создают сложную иерархическую сеть понятий, и знание о том, где находится понятие в этой сети, является важной частью знания об этом понятии. Свойства отношений различны при описании различных частей речи.

В разных системах тезаурус может выполнять разные функции:

- источник специальных знаний в узкой или широкой предметной области, способ описания и упорядочения терминологии предметной области;
- инструмент поиска в информационно-поисковых системах;
- инструмент ручного индексирования документов в информационно-поисковых системах;
- инструмент автоматического индексирования текстов [5].

Классическое толкование определения тезаурус в научной литературе состоит в определении его как словаря особого типа, отражающего словарный состав языка в полном объёме. Структура разработанного тезауруса подробно описывается в работах Ю. Н. Караулова [1; 2].

1. Методы анализа текстов

Первоначально, прежде чем говорить о методах формирования тезаурусов необходимо понять, что послужило началом создания тезауруса автоматизированно.

Первое использование автоматизации обработки текстов - это машинный перевод. Стали создаваться системы, в которых были проведены попытки создать автоматизацию формирования тезауруса, то есть отдать роль выбора метода, параметров формирования тезауруса информационным интеллектуальным системам.

В это время начали формироваться первые области исследований, которые в последующем оказали заметное влияние на возникновение научного направления, получившего название искусственный интеллект (ИИ).

Всю совокупность представленных на сегодняшний день методов анализа текста можно разделить на две большие группы:

- лингвистический анализ;
- статистический анализ.

Первый ориентирован на извлечении смысла текста по его семантической структуре. Второй - по частотному распределению слов в тексте. Деление на группы условное, так как в реальных задачах и при решении проблем всегда используется сочетание методов для достижения того или иного результата.

Для представления методов и моделей автоматизированного синтеза тезаурусов для начала ознакомимся с лингвистическим и статистическим анализом текстов.

1.1. Лингвистический анализ

Лингвистический анализ можно разделить на четыре взаимодополняющих анализа.

- *Лексический анализ.* Заключатся в разборе текстовой информации на отдельные абзацы, предложения, слова, определении национального языка изложения, типа предложения, выявлении типа лексических выражений (бранных, жаргонных слов) и т.д. Данный вид анализа не представляет существенной сложности для реализации.

- *Морфологический анализ*. Сводится к автоматическому распознаванию частей речи каждого слова текста (каждому слову ставится в соответствие лексико-грамматический класс). Часто морфологический анализ используется в статистических методах анализа при предварительной процедуре обработки документов - приведение слов к базовой форме.

Морфологический анализ для русского языка можно реализовать практически со 100% точностью благодаря его развитой морфологии. Для английского языка алгоритмы, присваивающие каждому слову в тексте наиболее вероятный для данного слова лексико-грамматический класс (синтаксическую часть речи), работают с точностью около 90%, что обусловлено лексической многозначностью английского языка.

- *Синтаксический анализ.* Заключатся в автоматическом выделении семантических элементов предложения - именных групп, терминологических целых, предикативных основ. Это позволяет повысить интеллектуальность процесса обработки тестовой информации на основе обеспечения работы с более обобщенными семантическими элементами.

- *Семантический анализ*. Заключатся в определении информативности текстовой информации и выделении информационно-логической основы текста. Проведение автоматизированного семантического анализа текста предполагает решение задачи выявления и оценки смыслового содержания текста. Данная задача является трудно формализуемой вследствие необходимости создания совершенного аппарата экспертной оценки качества информации.

Само создание экспертной системы уже говорит о том, что присутствует некоторый «интеллект», который помогает анализировать текст. Реализация семантического анализа текстовой информации предполагает обязательное использование экспертных систем, систем искусственного интеллекта для выявления смыслового содержания информации. В настоящее время отсутствуют сложившиеся подходы к реализации задачи семантического анализа текстовой информации, что во многом обусловлено исключительной сложностью проблемы и недостаточно полной проработкой научного направления создания систем искусственного интеллекта [4].

1.2. Статистический анализ

Статистический анализ - это, как правило, частотный анализ в тех или иных его вариациях. Общая суть такого анализа заключается в подсчете количества повторений слов в тексте и использовании результатов подсчета для конкретных целей (вычисление весовых коэффициентов ключевых слов) [Там же].

Всевозможные варианты различных реализаций подсчета слов и последующая обработка результатов подсчета образуют широкий спектр предлагаемых в данном классе методов и алгоритмов.

1.3. Латентно-семантический анализ

Латентно-семантический анализ (LSA - Latent Semantic Analysis) - это теория и метод для извлечения контекстно-зависимых значений слов при помощи статистической обработки больших наборов текстовых данных. Данный метод анализа используется не только в области поиска информации, но также в задачах фильтрации и классификации.

Основная идея латентно-семантического анализа заключается в том, что совокупность всех контекстов, в которых встречается и не встречается данное слово, задает множество обоюдных ограничений, которые позволяют определить похожесть смысловых значений слов и множеств слов между собой.

Исходной информацией для *LSA* является матрица термов на документы, которая описывает используемый для обучения системы набор данных. Элементы этой матрицы содержат частоты использования каждого терма в каждом документе [Там же].

Рассмотрим один из наиболее эффективных статистических подходов.

2. Методы анализа, позволяющие сформировать тезаурус

Обычно выделяют две группы методов анализа текстов, на основе которых строится тезаурус. К таким группам методов относятся: количественные и логические. В результате анализа текста осуществляется оценка в виде некоторых показателей.

В количественных (статистических) методах анализа каждой оценке ставится в соответствие ее количественная характеристика, полученная путем измерения с использованием числовой шкалы. Количественный анализ знаний приводит к получению обобщенных количественных характеристик - статистик. В зависимости от того, как получена статистика, количественный анализ дает точное числовое значение, приближенное, или вероятностное. Множество количественных характеристик конкретного знания является его количественной моделью.

В **логических методах анализа** используются номинальные шкалы, и каждой оценке ставится в соответствие некоторое высказывание. Логический анализ знаний приводит к получению логических выражений об истинности или ложности знания. В зависимости от того, как осуществляется анализ, логические методы разделяются на диалектические и формально-логические. Формально-логическое описание знания является его логической моделью и осуществляется на основе формального языка, формальной системы [3].

2.1. Статистические методы

Первичным материалом в лингвистической статистике является текст, рассматриваемый как последовательность лингвистических единиц заданного уровня: букв или фонем, морфов или морфем, словоформ или лексем, словосочетаний, предложений. На этом материале изучаются количественные характеристики лингвистических форм - их употребительность, совместная встречаемость, законы распределения в тексте, их физические размеры. На основе полученных данных описываются свойства текста, формулируются гипотезы о механизмах его образования и об устройстве системы языка. Основные понятия и категории в лингвистической статистике заимствуются у математической статистики. Такими понятиями являются понятия генеральной совокупности и выборки, частоты и вероятности, вероятностные распределения и статистические оценки. Однако применение этих понятий к лингвистическому материалу имеет ряд особенностей. В частности, в языкознании могут быть рассмотрены два принципиально разных вида генеральной совокупности: либо совокупность текстов одинакового жанра, заданного списка авторов или заданного временного интервала, либо совокупность единиц, принадлежащих одному лингвистическому уровню: фонем, морфем, слов или предложений [Там же].

Методы носят название статистические в силу того, что в их основе лежит подсчет статистики по законам Ципфа и Мандельброта.

Исследования количественных параметров текстов проводятся достаточно давно. Наиболее известные результаты были получены Ж.-Б. Эстопом (Jean Baptiste Estoup) и связаны с эмпирическим анализом использования слов в естественно-языковых текстах. Первым теоретическим результатом в области статистического анализа текста считается эмпирический закон, установленный Дж. К. Ципфом (George Kingsley Zipf) и получивший название «закона частот слов». Закон связывает гиперболической зависимостью частоту встречаемости слова в тексте с рангом этого слова в списке, упорядоченном по убыванию частот:

$$i(k, r) = pk\, r^{-b} \quad (1)$$

где i(k,r) - частота слова в тексте, k - общее число слов в тексте, r - ранг слова, т.е. его порядковый номер в упорядоченном по убыванию частотной функции словнике [Там же].

Статистические методы автоматизированного построения тезаурусов ПО легко представить алгоритмом и, соответственно, реализовать для автоматического построения тезауруса.

2.2. Логико-статистические методы

Методы имеет такую же структуру, как и статистические. Но здесь используется частотная информация, которая позволяет получить количественную характеристику связности понятий в тексте.

2.2.1. Дистрибутивно-статистический метод

Дистрибутивно-статический метод позволяет на основе частотной информации получать по некоторой заданной формуле количественную характеристику их связанности. Философия данного метода состоит в том, «что семантическую классификацию значимых элементов языка можно с большим основанием индуктивно извлечь из анализа текста, чем получить ее с некоторой точки зрения, внешней по отношению к структуре языка. Следует ожидать, что такая классификация даст более надежные ответы на проблемы синонимии и выражения смысла, чем существующие тезаурусы и списки синонимов, основанные главным образом на интуитивных ощущаемых сходствах без адекватной эмпирической проверки». В основе всех вариантов метода лежат количественные оценки, которые характеризуют совместную встречаемость языковых единиц текста в контекстах определенной величины. Основная гипотеза метода состоит в том, что слова, встречающиеся вместе в пределах некоторого текстового интервала, как-то связаны между собой. Для оценки связанности вводится коэффициент «силы связи», который рассчитывается по некоторой формуле. Вне зависимости от вида формулы в ней обычно используются характеристики совместной встречаемости пар слов и одиночной встречаемости каждого из слов [Там же].

2.2.2. Компонентный анализ

Метод компонентного анализа позволяет установить связь между двумя понятиями на основе анализа их дефиниций. Для реализации метода необходимым является наличие словаря определений.

Дефиниция - логическая операция:

1) раскрывающая содержание (смысл) имени посредством описания существенных и отличительных признаков предметов или явлений, обозначаемых данным именем (денотата имени);

2) эксплицирующая значение термина.

2.2.3. Частотно-семантический метод

Метод частотно-семантического анализа (ЧСА) является развитием метода компонентного анализа. Существо метода состоит в использовании в качестве критерия оценки семантической «силы связи» между словами одновременно двух характеристик дефиниций этих слов: общности дефинирующих элементов и частоты их встречаемости.

Метод предложен Ю. Н. Карауловым. С помощью такого метода был построен один из первых компьютерных семантических словарей русского языка [1; 2].

Исходными данными для ЧСА являлись: некоторые идеографические словари - они использовались для составления списка дескрипторов, краткий толковый словарь русского языка для иностранцев для составления списка слов, толковые словари С. И. Ожегова и Д. Н. Ушакова для установки дефиниций слов и дескрипторов [3].

В основе метода ЧСА лежит идея о целостности (интегрированности) ПО и отражении этого в ПО и, в частности, в языке. Образное представление идеи выражается следующей цитатой: «...представьте себе силы семантического притяжения в виде повсеместно существующего, разлитого в языке поля, в которое помещены тела - лексические единицы языка. Разные единицы в этом поле взаимодействуют между собой так же, как атомы, молекулы, макротела, планеты и космические объекты и на одном уровне, т.е. с однородными единицами, так и между уровнями» (взаимодействие происходит как на Рис. 1).

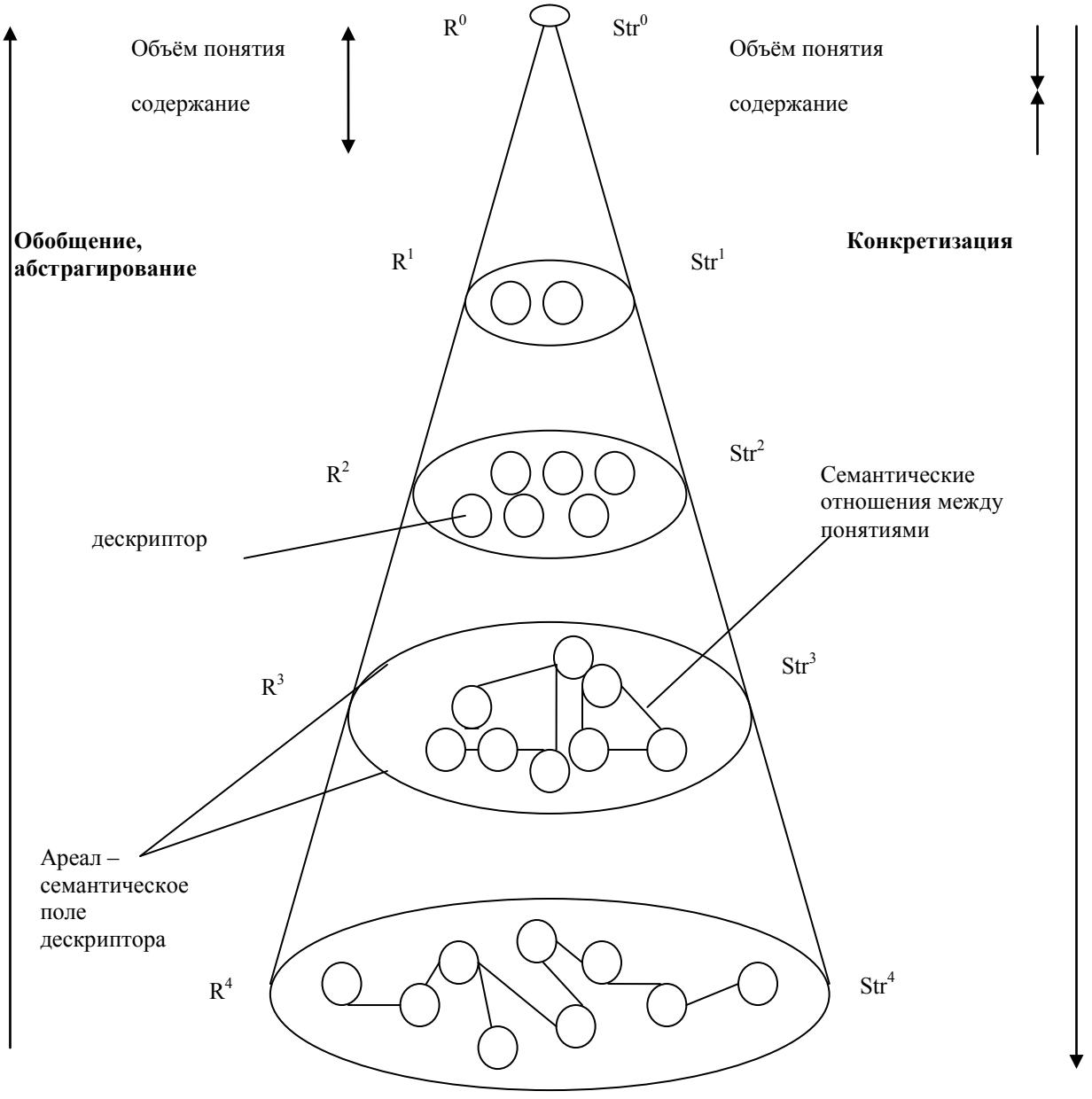

Рис. 1. *Представление понятий по уровням (по Str определяют уровень, а по R - ранг)*

Можно встретить точку зрения, что технология построения тезауруса целиком и полностью зависит от человека. Ю. А. Шрейдер отмечал, что «более перспективным методом является тот, на основании которого человек отбирает слова и значимые пары, указывая их смысловые сходства, а последующее разбиение материала на рубрики и сам выбор рубрик можно автоматизировать» [3; 5]. Легко заметить, что подобная технология имеет узкое место, связанное с отбором терминов и определением связей между ними. Понятно, что ввиду исключительной сложности задачи сегодняшние автоматизированные средства не позволяют полностью заменить человека, но это не значит, что невозможно получение максимально полезной для составления тезауруса информации исходя из анализа текстов.

3. Структура алгоритма автоматизированного синтеза тезаурусов на основе анализа текстов

Выделим основные направления действий с использованием методов анализа текстов для автоматизированного синтеза тезаурусов предметных областей (ПО):

- Для начала нужно выделить текст и, желательно, чтобы он соответствовал определенной ПО.
- Выбирается метод анализа текста в соответствии с тематикой и тем, что должен получиться тезаурус ПО, а не обычный словарь.
- На основании выбранного метода анализа текста анализируется текст ПО.
- В полученном варианте выделяется частота встречаемости слов, взаимосвязи слов. Это можно сделать исходя из двух законов: закона Ципфа и закона Мандельброта.
- Строится граф-модель текста, расставляются связи и проставляются частоты встречаемости.
- Убираются лишние слова: союзы, местоимения и т.д., т.е. те, которые не задействованы в связях, полученных после построения графовой модели.
- На основании большого числа встречаемости слов строятся отдельные предложения (понятия). Для этого выбираются из графовой модели отдельные подграфы.
- Для каждого из подграфов строится словарь - тезаурус ПО, т.е. от графовой модели переходят к тексту.
- Полученный тезаурус нужно проверить на принадлежность каждого определения ПО.

Заключение

В статье была рассмотрена тема модели и методы автоматизированного синтеза тезауруса предметных областей. Тема является важной не только в сфере обучения, но и для любой другой деятельности, для которой познание предмета или предметной области является очень важным аспектом.

Классическое толкование тезауруса в научной литературе состоит в определении его как словаря особого типа, отражающего словарный состав языка в полном объеме.

Понятие тезауруса может толковаться расширительно, если в качестве языковых единиц, представляемых в нем рассматривать предложения, сверхфразовые единства и тексты. В этом случае тезаурус будет являться своеобразным гипертекстом, т.е. естественно-языковым описанием.

В качестве элементов тезауруса можно рассматривать отдельные суждения, а в качестве семантических отношений выбрать логические операции, тогда тезаурус будет представлять собой некий свод возможных логических рассуждений, некоторую логическую модель.

Если соотносить автоматизированное формирование тезауруса с искусственным интеллектом, то есть создание тезауруса с минимальным участием человека, то можно выделить следующие моменты:

1. *Представление знаний*. В тезаурусе знания представляются в виде понятий и описываются в виде языка, понятного системам искусственного интеллекта.

2. *Общение*. Сформированный тезаурус или тезаурус, представленный на этапе формирования, должен иметь средства и способы общения, контроля, выраженные или заложенные в системах ИИ.

3. *Рассуждения*. Подобно тому, как человек принимает решение о вхождении того или иного термина в тезаурус, система ИИ должна уметь учитывать это, так же как и учитывать сложность понятий, разбивать их на более мелкие и не сложные к восприятию.

4. *Восприятие*. Программа или система ИИ должна правильно понимать и воспринимать данные (текст), которые поступают на автоматизированную переработку и дальнейшее формирование тезауруса предметных областей.

5. *Обучение*. ИИ должен обучаться в процессе деятельности по созданию тезаурусов, находить новые интегрированные методы формирования тезаурусов, создавать более точные и адекватные модели.

6. *Деятельность*. Она представляет собой осмысленное формирование тезауруса. ИИ может задуматься: «а зачем он формирует тезаурус?».

Всю совокупность представленных на сегодняшний день методов анализа текста можно разделить на две большие группы:

- лингвистический анализ;
- статистический анализ.

Существуют несколько методов автоматизированного построения тезауруса предметных областей:

- Статистические методы, которые используют закон Ципфа и закон Мандельброта.
- Дистрибутивно-статистический метод.
- Компонентный анализ.
- Частотно-семантический метод.

Рис. 2. *Структура алгоритма автоматизированного синтеза тезаурусов предметных областей*

Список литературы

1. **Караулов Ю. Н.** Лингвистическое конструирование и тезаурус литературного языка. М.: Наука, 1981. 366 с.
2. **Караулов Ю. Н.** Частотный словарь семантических множителей русского языка. М.: Наука, 1980. 207 с.
3. **Филиппович Ю. Н., Прохоров А. В.** Семантика информационных технологий: опыты словарно-тезаурусного описания / с предисловием А. И. Новикова. М.: Изд-во МГУП, 2002. 368 с.
4. **Чугреев В. Л.** Модель структурного представления текстовой информации и метод её тематического анализа на основе частотно-контекстной классификации: дисс. … к. техн. н. СПб.: ЛЕТИ, 2003. 185 с.
5. **Шрейдер Ю. А.** Информация в структурах с отношениями // Исследования по математической лингвистике, математической логике и информационным языкам. М.: Наука, 1972. С. 147-159.