Федеральное агентство по образованию Российской Федерации

Нижнетагильская государственная социально-педагогическая академия

Кафедра прикладной информатики

**ИНФОРМАЦИОННО-ПОИСКОВЫЙ ТЕЗАУРУС**

**Контрольная работа**

Исполнитель:

студент(ка) 531(д) ИФ (ОЗО)

Путяшева Е.С.

Проверил:

Гуторова Л.Е.

Нижний Тагил

2007

**Оглавление**

Введение

1. Информационные системы
2. Информационно-поисковый тезаурус: определение, цель разработки
3. Использование информационно-поисковых тезаурусов
4. Структура информационно-поискового тезауруса
5. Построение информационно-поисковых тезаурусов
6. Сфера применения и перспективы развития информационно-поисковых тезаурусов

Заключение

Литература

**Введение**

В настоящее время в связи с возрастанием объема документальной информации, необходимой для принятия эффективного управленческого решения, и возможностями, предоставляемыми новыми компьютерными технологиями, автоматизированные информационно-поисковые системы стали широко использоваться в различных сферах экономики.

Организация быстрого и эффективного поиска документальной информации становится все более неразрешимой проблемой.

Цель документального поиска - нахождение и выдача соответствующих запросу пользователя документов или их описаний.

Традиционно информационно-поисковые системы (ИПС) применяются для тематического поиска научно-технической информации в крупных библиотеках, научно-технических центрах, архивах, патентных библиотеках.

Таким образом, сфера приложения для технологий информационно-поисковых систем представляется достаточно широкой.

При автоматизации поиска документальной информации важнейшей является задача формализации содержания документа и запроса. При этом поиск происходит по всему тексту документа или по его поисковому образу, а в качестве запроса чаще всего выступают отдельные ключевые слова или их логические комбинации. Именно на этой технологии основано действие информационно-поискового тезауруса (ИПТ). Цель моей работы – дать необходимые определения, определить цели разработки ИПТ, его структуру и построение, а также значение и перспективы развития.

**1. Информационно-поисковые системы**

Автоматизированный документальный поиск может быть организован на основе различных технологий: поиска по поисковому образу документа, поиска по полному тексту документа, поиска документов по гипертекстовым ссылкам.

Технология полнотекстового поиска является неотъемлемой составляющей таких современных и перспективных информационных технологий, как: системы управления документами (Document management system, DMS), технологии групповой работы над документами (groupware), технологии поиска в Internet/intranet. На технологии гипертекста базируется самый известный сервис Internet World Wide Web (WWW).

Информационно-поисковая система для управленческих документов, как правило, требует раз­работки собственного информационно-поискового языка, адекватного данной предметной области.

В документальных информационно-поисковых системах поиск документа происходит по краткому формализованному описанию его содержания - так называемому поисковому образу документа.

Важнейшей структурной составляющей является информационно-поисковый язык.

Информационно-поисковый язык, ИПЯ (indexing language, retrieval language) - искусственный язык для выражения содержания документов или запросов с целью последующего поиска.

Основное назначение ИПЯ - установить принадлежность того или иного документа к определенной группе понятий.

Перевод текстов документов и запросов на ИПЯ называется индексированием.

В результате индексирования содержание документа отображается в поисковый образ документа, а содержание запроса - в поисковое предписание.

Индексирование может быть ручным (когда его производит человек) или автоматическим.

Для общения человека с компьютером разрабатываются специальные искусственные языки, естественный язык (ЕЯ) в этом качестве пока не может быть использован без специального предварительного преобразования. В естественном языке присутствуют: синонимия, омонимия и полисемия, избыточность, субъективность и другие свойства, препятствующие его автоматизированной обработке.

Синонимы - слова, различающиеся по написанию, но совпадающие по смыслу (Россия, РФ, Российская Федерация).

Омонимы - слова, совпадающие по звуковому и графическому составу, но имеющие разные значения (ключ может быть - телеграфный, поисковый, к замку, к шкафу и т.д.).

Полисемия - наличие у одного и того же слова нескольких разных, но связанных между собой значений (например: фамилия - как имя и фамилия - как семья).

Избыточность ЕЯ - наличие лишней информации, без которой можно точно и однозначно установить смысл или значение сообщения. Таким образом, в естественном языке часто нет однозначного соответствия между словом и его значением.

Искусственный язык, как правило, разрабатывается на основе ЕЯ. При этом устраняется многозначность слов ЕЯ.

Слова любого языка в процессе отображения предметов реального мира вступают между собой в определенные отношения. Эти отношения можно разделить на парадигматические и синтагматические.

Парадигматические отношения - логические отношения, существующие между лексическими единицами ИПЯ независимо от контекста, в котором эти лексические единицы употребляются. Эти отношения обусловлены предметно-логическими, а не языковыми факторами, т.е. относятся к категории внеязыковых связей. Примеры парадигматических отношений: часть - целое (отдел - организация); род - вид (ценная бумага - акция); причина - следствие; функциональное сходство; ассоциации. Учет парадигматических отношений необходим для правильного выбора и точного употребления слов.

Наиболее часто в качестве основания деления при классификации ИПЯ используют способ организации понятий. По способу организации понятий различают: предкоординируемые (классификационные) ИПЯ; посткоординируемые (дескрипторные) ИПЯ.

Дескрипторные языки семантически более сильны, чем классификационные, но более сложны для формальной обработки.

Дескриптор - ключевое слово или словосочетание, служащее именем класса эквивалентных ключевых слов (синонимов). Как правило, в качестве дескриптора выбирают одно, наиболее употребляемое ключевое слово из класса синонимичных ключевых слов.

Разработка дескрипторного языка фактически сводится к разработке информационно-поискового тезауруса.

1. **Информационно-поисковый тезаурус: определение, цель разработки**

Тезаурус – термин, широко используемый в информатике как составная часть информационно-поисковых систем.

Можно выделить два определения информационно-поискового тезауруса:

1. Информационно-поисковый тезаурус представляет собой словарь, отображающий семантические отношения между лексическими единицами дескрипторного информационно-поискового языка (дескрипторами) и предназначенный для поиска слов по их смысловому содержанию.

2. Информационно-поисковый тезаурус (ИПТ) – контролируемый словарь терминов предметной области, создаваемый для улучшения качества информационного поиска в данной предметной области.

И в том, и в другом случае речь идет о словаре, который призван облегчить поиск необходимой информации.

Возможно два способа расположения слов в словарях: по близости их буквенного состава и по смысловой близости.

По первому способу создаются алфавитные словари. По второму способу - тезаурусы. Алфавитные словари служат для раскрытия значения данного слова. Тезаурусы служат для поиска слов для выражения данного понятия. Т.е., если в обычном словаре по слову ищется его смысл, то в тезаурусе по заданному смыслу ищутся слова, которые этот смысл выражают.

Цели разработки ИПТ:

- перевод языка авторов на контролируемый язык, используемый для индексации и поиска;

- обеспечение последовательности в присваивании индексных терминов;

- обозначение отношений между терминами;

- облегчение информационного поиска.

**3. Использование информационно-поисковых тезаурусов**

Тезаурусы используют в своей работе специалисты, в той или иной степени связанные с подготовкой и обработкой текстовой информации: переводчики, редакторы, научные работники, референты и др.

Информационно-поисковые тезаурусы используются при индексировании документов и запросов.

Различают безтезаурусное индексирование и индексирование с тезаурусным контролем лексики.

При безтезаурусном индексировании поисковый образ документа представляется в виде списка ключевых слов, выбранных непосредственно из текста документа. Такое индексирование требует меньше времени и менее трудоемко. Однако такой режим индексирования обладает рядом недостатков: возможна синонимия; нельзя отразить парадигматические отношения.

При индексировании с тезаурусным контролем лексики для выражения содержания документа и запроса используют только дескрипторы, входящие в информационно-поисковый тезаурус, что позволяет исключить синонимию и учесть парадигматические связи, и, в конечном счете, повышает эффективность поиска.

Таким образом, информационно-поисковые тезаурусы используются для:

- перевода текста с ЕЯ на дескрипторный ИПЯ путем замены ключевых слов соответствующими дескрипторами, при этом исключается многозначность;

- отображения парадигматических связей между дескрипторами.

**4. Структура информационно-поискового тезауруса**

Тезаурус имеет вид словаря, в котором в определенном порядке расположены термины и зафиксированы связи между ними.

Как правило, тезаурус состоит из введения, алфавитного указателя (словарная часть) и систематического указателя.

Введение содержит общие характеристики тезауруса - область применения, правила использования, число дескрипторов, ключевых слов и другую справочную информацию.

Алфавитный указатель включает алфавитный перечень дескрипторных статей.

Дескрипторная статья может иметь следующую структуру:

- заглавный дескриптор;

- ключевые слова из класса эквивалентности;

- дескрипторы, подчиняющие заглавный;

- дескрипторы, подчиненные заглавному;

- дескрипторы, ассоциированные с заглавным.

Ассоциации могут быть следующих видов: причина - следствие; сырье - продукт; процесс - объект; процесс - субъект; свойство - носитель свойства; функциональное сходство и др.

Систематический указатель служит для раскрытия, учета и контроля парадигматических отношений между дескрипторами. Для каждого дескриптора ИПЯ должны быть найдены соответствующие родовые и видовые дескрипторы. Причем предварительно необходимо сформулировать критерии, определяющие число уровней иерархии. Определение степени отображения парадигматических отношений зависит от требований, предъявляемых к системе. Чем шире эти отношения отображены, тем выше полнота выдачи, но одновременно происходит увеличение словарного запаса, усложняется система ссылок и может произойти уменьшение точности поиска.

**5. Построение информационно-поисковых тезаурусов**

Тезаурусы разрабатываются для отдельных отраслей. Например, существует «Тезаурус по документации», «Дескрипторный словарь по информатике» и т.п. Тезаурус может быть разработан и для обслуживания конкретной информационно-поисковой системы.

Построение тезауруса состоит из нескольких взаимосвязанных этапов.

Первый этап - формирование словника. Словник - первоначальные множества ключевых слов. При этом рассматривается представительный массив наиболее информативных для данной предметной области документов. Например: реферативные журналы, словари, учебники, справочники, нормативные документы и т.д. Выбираются слова, употребляемые в этих источниках, при этом устанавливается частота употребления слов и учитываются все формы, которые могут иметь слова.

Второй этап - формирование множества ключевых слов. Из словника формируется множество ключевых слов. При отборе ключевых слов учитывается информативность слова, которая определяется исходя из частоты встречаемости слова, роли слова в данной предметной области. Процесс выбора ключевых слов достаточно сложно формализовать. Например, такой критерий, как частота встречаемости не может быть абсолютным. Если слово встречается в текстах очень часто, это может означать, что оно выражает чрезмерно широкое понятие, либо недостаточно четко определено, т.е. неинформативно. Если ключевое слово встречается очень редко, это может означать, что оно выражает новое понятие и таким образом является информативным.

Третий этап - формирование классов эквивалентности. Выделение дескрипторов.

Класс эквивалентности - это группа терминов, равнозначных по смыслу в пределах сферы действия данной системы. Одно из слов класса эквивалентности объявляется дескриптором (как правило, наиболее употребляемое). Все слова, включенные в класс эквивалентности, объявляются условными синонимами (условными - так как они равнозначны в пределах данной предметной области, а в ЕЯ эти слова могут не являться синонимами).

Критерием включения слов в класс эквивалентности является семантическая значимость этого слова при поиске документов, т.е. если данное слово при поиске документа может быть заменено другим словом, так, что на любой запрос выдача документов будет такой же, как и до замены, то такие два слова объявляются условными эквивалентами и включаются в один класс эквивалентности. Ключевые слова, относящиеся к одному классу эквивалентности, помещаются в соответствующую дескрипторную статью и помечаются символом «с».

В процессе построения тезауруса и выделения множества дескрипторов происходит устранение синонимии, омонимии, полисемии ключевых слов, а также выявление парадигматических связей между дескрипторами.

В ЕЯ многие слова могут быть правильно истолкованы только с учетом контекста. В ИПЯ для правильной трактовки ключевых слов необходимы средства для исключения многозначности трактования слов.

Для устранения многозначности (омонимии и полисемии) ключевое слово, которое можно неоднозначно трактовать, снабжается пометой, в которой указывается в каком значении данное ключевое слово используется в ИПЯ.

Разработка тезауруса без использования компьютерных технологий - достаточно длительный и трудоемкий процесс.

Большинство автоматизированных методов создания тезауруса основано на автоматическом подсчете частоты слов, содержащихся в документах, считающихся типичными для данной предметной области. Результаты такой обработки обычно представляются в виде матрицы «документ-термин», (рис.1). Элемент матрицы на пересечении строки и столбца, трактуется как вес слова в документе. На основе полученной матрицы вычисляются коэффициенты подобия между словами, определяемые в зависимости от частоты, с которой слова совместно встречаются в анализируемых документах, и формируется структура тезауруса.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Документы | Частота слова в документе | | | | | | |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| D1 D2 D3 D4 |  |  |  |  |  |  |  |

Рис.1. Пример матрицы «термин-документ»

Автоматические информационно-поисковые тезаурусы являются составным элементом систем автоматического индексирования документов и запросов. В словарной статье автоматического тезауруса, как правило, зафиксированы отношения условной эквивалентности (синонимии), отношение подчинения и ассоциативные отношения.

**6. Сфера применения и перспективы развития информационно-поисковых тезаурусов**

В настоящее время происходит расширение сфер применения автоматических тезаурусов. При этом тезаурусы выступают составной частью современных систем подготовки текстов, осуществляя лингвистическую поддержку процесса подготовки и обработки текстов на естественном языке.

Среди наиболее перспективных направлений развития автоматических тезаурусов можно указать следующие:

1. Получение справки по используемому слову. Указав слово, в качестве ключа для запроса, пользователь в ответ получает соответствующий фрагмент словаря, содержащий лингвистическую информацию о данном слове. Например, автоматический тезаурус получает от пользователя некоторое существительное и в ответ выдает совокупность устойчиво сочетающихся с ним глаголов или все наиболее часто сопровождающие его определения. При этом автоматически выполняется процедура нормализации входного слова (т.е. приведение существительного к именительному падежу).

2. Контекстные замены по требованию пользователей. В данном случае тезаурус не только подбирает вместо одного словосочетания другое, которое пользователь счел более соответствующим контексту по смысловым или стилистическим соображениям, но и автоматически переоформляет параметры слов (например, род прилагательного) в соответствии с контекстом. Это означает, что синтаксические операции, производимые тезаурусом, существенно усложняются.

3. Автоматическая оценка стиля. Если слова и словосочетания в тезаурусе снабдить стилистическими пометками, то он может использоваться для стилистической оценки текста с выделением слов и словосочетаний, выпадающих, из общего стиля документа.

**Заключение**

В условиях конкурентной борьбы, необходимости постоянного повышения качества работы органа управления и всего предприятия в целом на руководителя обрушивается мощный поток информации, прежде всего документальной, и у него возникают проблемы с ее переработкой. Необходимы специальные методы информационной обработки документов, организации быстрого и эффективного поиска документальной информации. Именно информационно-поисковые тезаурусы служат решению этой проблемы.

Считаю, что в моей работе достаточно подробно и логично изложены способы использования и построения информационно-поискового тезауруса, а также сфера его применения и перспективы развития.

**Литература**

1. Ивасенко А.Г. Информационные технологии в экономике и управлении: учебное пособие / А.Г.Ивасенко, А.Ю.Гридасов, В.А.Павленко.- 2-е изд., стер.- М.: КНОРУС, 2007.- 160с.

2. Информатика для юристов и экономистов / Симонович С.В. и др.- СПб.: Питер, 2001.- 688с.: ил.

3. Информатика: Базовый курс/С.В.Симонович и др.- СПб.: Питер, 2002.- 640с.: ил.

4. Михеева Е.В. Информационные технологии в профессиональной деятельности: учебное пособие.- М.: ТК ВЕЛБИ, Изд-во Проспект, 2007.- 448с.

5. Организация работы с документами: Учебник / Под ред.проф.В.А.Кудряева.- 2-е изд., перераб. и доп.- М.: ИНФРА-М, 2001.- 592с.- (Серия «Высшее образование»).