МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Саратовский национальный исследовательский
Государственный
университет имени Н. Г. Чернышевского»

Геологический колледж СГУ

Исследовательский проект

на тему: «Дескрипторные информационно-поисковые языки»

Выполнил: студент группы 2132

Баннов Е.С.

Руководитель: преподаватель информатики

Жумушева М.М.

Саратов

2023

Содержание

Введение

В современном мире огромную роль в жизни людей играет информация. Постоянная, регулярная работа с информацией в наше время стала неотъемлемой частью жизни каждого цивилизованного современного человека. Человеку, в силу своей профессии или увлечений часто сталкивающемуся с подбором и поиском какой-либо тематической информации, рано или поздно (с возрастанием ее объема) приходится применять некоторые принципы систематизации и классификации имеющихся данных, обеспечивающие более удобный и эффективный поиск.

Гипотеза исследования: я предполагаю, что с усовершенствованием старых и созданием новых языков программирования людям стало легче работать, а программирование стало не таким сложным.

Цель исследования: изучить информацию о Дескрипторных информационно-поисковых языках, выписать их основные понятия.

Задачи:

1. Узнать виды, классификации.
2. Исследовать сферы применения языков.
3. Разобрать основные понятия.
4. Сделать выводы о проделанной работе.

Объектом исследования, является: дескрипторные информационно-поисковые языки.

Методы исследования:

1. Работа с информационно-поисковыми языками.
2. Обобщение. Практическая значимость: с помощью данного проекта каждый человек сможет ознакомиться с дескрипторными информационно-поисковыми языками.

Глава 1 . Основные понятия дескрипторных-поисковых языков и их свойства.

1.1. Основные понятия.

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 managementsystem, DMS*), технологии групповой работы над документами (*groupware*), технологии поиска в *Internet/intranet*. На технологии гипертекста базируется самый известный сервис *Internet World Wide Web (WWW)*.

Первоначальным направлением развития СУБД стала разработка и использование фактографических информационных систем, которые ориентированы на обработку структурированных данных. Были разработаны модели организации фактографических данных, отработаны программно-технические решения по накоплению и физическому хранению таких данных, реализованы языки запросов к БД.

Однако создание фактографических информационных систем требует предварительной структуризации данных, например, на основе *таблиц*. Она зачастую требует больших накладных расходов. Вместе с тем накапливаются большие объемы неструктурированной информации: в организационно-распорядительных документах или других текстовых источниках. Представление такой информации в фактографических системах зачастую экономически не оправдано. Теоретические исследования вопросов автоматизации обработки неструктурированной информации, начавшись еще в 50-х годах, пока не привели к созданию такой строгой, полной и технически реализуемой модели представления и обработки данных, как реляционная модель. Пока не разработаны стандартные информационно-поисковые языки

(подобные SQL), которые можно было бы использовать для формализованного описания содержания документов и построения запросов.

Элементом данных в документальных ИС является документ (в фактографических информационных системах элементом является запись). Обычно под документом понимается текстовый файл.

Основной задачей документальных информационных систем является хранение и предоставление пользователю документов, содержание которых соответствуют его информационным потребностям.

Документальная информационная система (ДИС) — единое хранилище документов с инструментарием поиска и выдачи необходимых пользователю документов.

Поисковый характер документальных информационных систем (определил еще одно их название — информационно-поисковые системы (ИПС).

Соответствие найденных документов информационным потребностям пользователя называется пертинентностью.

В зависимости от особенностей реализации хранилища документов и механизмов поиска, ДИС можно разделить на две группы:

» системы на основе индексирования;

» семантически-навигационные системы.

Семантика — значения единиц языка.

В семантически-навигационных (гипертекстовых) системах документы, помещаемые в хранилище документов, оснащаются специальными навигационными конструкциями (гиперссылками), соответствующими смысловым связям между различными документами или отдельными фрагментами одного документа.

В системах на основе индексирования исходные документы помещаются в базу без какого-либо дополнительного преобразования, но при этом смысловое содержание каждого документа отображается в некоторое поисковое пространство. Процесс отображения документа в поисковое

пространство называется индексированием и заключается в присвоении каждому документу некоторого индекса — координаты в поисковом пространстве. Формализованное представление индекса документа называется поисковым образом документа (ПОД). Пользователь выражает свои информационные потребности посредством специального языка, формируя поисковый образ запроса (ПОЗ) к базе документов.

На основе определенных критериев ДИС осуществляет поиск и выдачу документов, поисковые образы которых соответствуют поисковым образам запроса пользователя.

Соответствие найденных документов запросу пользователя называется релевантностью.

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

Информационно-поисковый язык (ИПЯ) представляет собой некоторую формализованную семантическую систему, предназначенную для выражения содержания документа и поискового запроса.

1.2. Свойства информационных поисковых языков.

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

ИПЯ состоит из алфавита, лексики и грамматики. **Алфавит** — система знаков, используемая для записи слов. В ИПЯ могут быть использованы: буквы латинского алфавита; кириллица; цифры; пунктуационные знаки. Лексика (словарный состав) — совокупность слов, входящих в состав языка, называемых также лексическими единицами. Лексическая единица — слово или семантически неделимое словосочетание, выражающее какое-либо понятие. **Грамматика** — набор правил, по которым из конечного числа элементов определенного типа (например, букв или слов) можно получить язык для выражения содержания документов или запросов или описания фактов с целью последующего поиска. Грамматика подразделяется на

морфологию и синтаксис. **Морфология** — правила построения и изменения слов. **Синтаксис** — правила построения и изменения соединения слов (построение фраз). 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. Поэтому в семантически развитом ИПЯ должны быть в явном виде выражены важнейшие отношения между терминами, иначе при отображении текста документа может произойти потеря или искажение смысла документа. Например, при поиске нормативных документов, касающихся термина «акция», для увеличения полноты поиска возможно указание термина «ценная бумага».

Синтагматические отношения — отношения слов при соединении их в словосочетания и фразы. Линейные логические отношения, которые устанавливаются между словами непосредственно при их использовании в тексте, объединяют эти слова в сочетания и предложения. Для уточнения .смысла документа или запроса, помимо ключевых слов, часто необходимо указывать в каких синтагматических отношениях эти слова находятся. Так, фраза «защита окружающей среды от человека» и фраза «защита человека от окружающей среды» имеют совершенно разный смысл/хотя и состоят из одних и тех же ключевых слов. Таким образом,

развитый ИПЯ должен обладать средствами отображения парадигматических и синтагматических отношений.

Для оценки сравнительной эффективности различных языков используется понятие *семантическая сила языка*.

Семантическая сила ИПЯ характеризует смысловыразительные возможности ИПЯ и показывает, насколько ИПЯ уступает ЕЯ. Семантическая сила тем больше, чем богаче словарный состав ИПЯ и шире его словообразовательные возможности (создание новых слов, соответствующих новым понятиям); шире используются средства отображения парадигматических и синтагматических отношений между словами.

Можно указать следующие требования, которым должен удовлетворять семантически развитый ИПЯ:

- располагать лексико-грамматическими средствами для точного отображения центральной темы документа и запроса;
- не содержать полисемии, синонимии и омонимии, т.е. каждая запись на ИПЯ должна допускать только одно толкование;
- отображать только объективные характеристики предметов и отношений между ними;
- быть удобным для алгоритмического сопоставления (отождествления) поискового образа документа (ПОД) и поискового предписания (ПП).

Как правило, чем больше семантическая сила ИПЯ, тем труднее с ним работать. Наиболее часто в качестве основания деления при классификации ИПЯ используют способ организации понятий.

По способу организации понятий различают:

- предкоординируемые (классификационные) ИПЯ;
- посткоординируемые (дескрипторные) ИПЯ.

Предкоординация — предварительное (до использования при индексировании) построение сложных классов путем логического умножения (координации) простых классов. Словарный состав задается в виде

фиксированного списка слов, словосочетаний и фраз. При индексировании документов или запросов можно пользоваться только словами, словосочетаниями и фразами, содержащимися в фиксированном списке. Введение в язык новых лексических единиц строго ограничено и возможно лишь до индексирования документов, т.е. при создании языка. Словарный состав предкоординируемых языков напоминает двуязычный разговорник, в котором заранее зафиксированы наиболее употребительные фразы. При помощи предкоординируемого языка происходит отнесение документа к классу, обозначенному лексическими единицами этого языка, т.е. классификация документа.

Посткоординируемые (дескрипторные языки) основаны на методе координатного индексирования. В посткоординируемых ИПЯ лексические единицы объединяются в поисковом образе лишь во время индексирования документа. Словарь дескрипторного 'ИПЯ состоит из специальным образом выбранных отдельных слов или словосочетаний ЕЯ — ключевых слов и дескрипторов.

Координатное индексирование — индексирование, при котором основное смысловое содержание текста (документа) или информационного запроса представляется в виде сочетания ключевых слов или дескрипторов.

Ключевые слова- это наиболее существенные для отображения содержания документа слова и словосочетания, обладающие назывной функцией.

Назывные слова - слова, обозначающие вещи, явления, процессы, имена собственные (т.е. в качестве ключевого слова не может выступать предлог, союз и др.)

Глава 2 . Информационные-поисковые языки и их классификации.

2.1.Основные информационные-поисковые языки.

Информационно-поисковые каталоги, основанные на классификации сведений по определенной предметной области, были первыми системами информационного поиска документов.

Классификация — это группировка объектов по признакам.

По области или по сфере применения информационно-поисковых языков можно **выделить**:

1. *Коммуникативные (общесистемные) ИПЯ* - предназначенные для обеспечения взаимодействия между различными (информационными, библиотечными и др.) системами (в том числе распределенными по государственной, ведомственной или территориальной принадлежности);
2. *Локальные (внутренние) ИПЯ* - предназначенные для использования в рамках отдельной системы;
3. *Внешние ИПЯ* - используемые в других системах и предназначенные для взаимодействия только с ними.

Различают языки описания (декларативные языки), которые в свою очередь подразделяются на языки предкоординатные (классификационные) и посткоординатные (дескрипторные), а также процедурные языки - языки запросов и манипулирования данными.

2.2.Классификационные информационно – поисковые языки.

К классификационным языкам относят:

- информационно-поисковый язык иерархического типа;
- информационно-поисковый язык фасетного типа;
- алфавитно-предметную классификацию.

Иерархическая классификация — это перечислительная классификация (т.е. все возможные классы заранее перечислены), в которой каждый класс делится на подклассы. Термины в иерархической классификации

расположены в порядке их перехода от общих понятий к частному. Классификация осуществляется в зависимости от выбранных оснований деления и порядка их следования. В иерархической классификации необходимо иметь отдельные исчерпывающие классы для всех возможных предметов, т.е. все возможные классы должны быть заранее перечислены, поэтому иерархическую классификацию и называют перечислительной.

Процедура построения ИПЯ иерархического типа включает следующие этапы:

- **Анализ предметной области**, определение оснований деления (признаков классификации). В качестве признаков класификации выбирают такие, по которым имеет смысл производить поиск документов в данной предметной области.
- **Установление соподчиненное признаков**. Соподчиненность может быть естественной или установленной.
- **Формирование классов документов на основе выбранных признаков классификации.** Получение иерархического дерева классов.
- **Формирование индексов каждого класса.**
- **Составление классификационных таблиц и алфавитного указателя**. В классификационной таблице классы упорядочены по индексу, а в алфавитном указателе - по алфавиту.

Индексирование с использованием ИПЯ иерархического типа заключается в определении того, к какому классу относится описываемый объект, и в определении по классификационной таблице и алфавитному указателю индекса этого класса. Преимущество языков иерархического типа состоит в простоте индексирования и поиска.

Классификация наиболее эффективна в том случае, когда классы в иерархической системе располагаются в естественном порядке и набор классов в течение времени не изменяется (т.е. предметы естественно находятся в жесткой иерархической соподчиненности). Например, классификация документов в организации, имеющей стабильную структуру.

ИПЯ фасетного типа основаны на принципах многоаспектной классификации, в которой каждый конкретный класс строится при индексировании по определенным правилам из предварительно заданных категориальных классов — **фасетов**. В системах фасетной классификации не ставится задача перечислить все сложные классы. Такие системы предлагают составные элементы, из которых по фасетной формуле составляется индекс.

Процедура разработки ИПЯ фасетного типа состоит из следующих этапов:

- Анализ предметной области, для которой составляется классификация. Выделение основных признаков классификации. Эти категории называются фасетами, которые при необходимости более детальной классификации могут делиться на субфасеты и т.д.
- Все возможные простые классы группируются по фасетам. Каждый простой класс фасета называется фокусом.
- Обозначение соответствующими шифрами фасетов и фокусов
- Установление фиксированной последовательности фасетов в поисковом образе (фасетная формула).
- Составление алфавитного указателя фасет и фокусов. Преимущество ИПЯ фасетного типа по сравнению с ИПЯ иерархического типа состоит в том, что допускается многоаспектное индексирование, так как существует возможность строить классы из разных сочетаний фокусов и получать любые сочетания заранее выбранных характеристик объектов классификации. На практике иерархическая и фасетная классификация часто используются в сочетании. Например, **УДК** — универсальная десятичная классификация. **Алфавитно-предметная классификация** — система классов, каждый из которых соответствует определенной теме или одному виду предметов, причем классы расположены в алфавитном порядке имен этих классов. Основной словарный состав (лексика) ИПЯ состоит из упорядоченных по алфавиту множества слов, словосочетаний и фраз ЕЯ.

2. Алфавитно-предметная классификация – это система классов, соответствующих определенной теме и расположенных в алфавитном порядке имен этих классов.

Алфавитно-предметная классификация содержит:

- **предметный заголовок** — слово, словосочетание или фраза ЕЯ, используемое для обозначения предмета или темы, заголовок может подразделяться на подзаголовки;
- **предметный словник (лексический состав языка)** — упорядоченное по алфавиту множество предметных заголовков, используемых для построения алфавитно-предметной классификации;
- **предметную рубрику** — совокупность предметного заголовка с описанием адреса хранения документов, основная тема которых обозначается этим предметным заголовком.

Алфавитно-предметная классификация предназначена для построения каталогов для узко предметного поиска. В таких каталогах под предметными заголовками даются сведения (шифр или библиографическое описание) документов, предмет которых обозначен данным заголовком.

Порядок составления алфавитно-предметной классификации:

- Анализ предметной области и выбор тем классификации.
- Устранение синонимии слов, словосочетаний и фраз, используемых в качестве предметного заголовка. В случае синонимии можно использовать систему ссылок.
- Выделение основных, ведущих слов в словосочетаниях и фразах, используемых в качестве предметных заголовков. Обозначение парадигматических связей между названиями предметов и тем. Эти связи обозначаются с помощью ссылок.

Алфавитно-предметная классификация используется главным образом для информационного поиска по отдельным предметам и темам. И применяется в качестве предметных указателей к каталогам документов.

Основной недостаток классификационных языков состоит в том, что они не обеспечивают возможности поиска документов по любому, заранее не заданному сочетанию признаков.

Глава 3 . Основа построения дескрипторных информационно-поисковых языков.

3.1.Построение дескрипторных языков.

В основе построения дескрипторных информационно-поисковых языков лежит принцип координатного индексирования, который предполагает, что основное смысловое содержание документа может быть выражено списком ключевых слов. К ключевым словам относятся так называемые полнозначные слова — *существительные, прилагательные, глаголы, наречия, числительные, местоимения*. Ключевыми словами не могут быть предлоги, союзы, связки, частицы.

Основными элементами ДИПЯ являются:

1. словарь лексических единиц;
2. правила применения ИПЯ (грамматика), определяющие процедуру перевода текстов документов и запросов с естественного языка на ИПЯ;
3. правила построения ИПЯ.

Словари лексических единиц делятся на две группы:

1. основные лексические словари, составляющие лексику ИПЯ;
2. морфологические словари, обеспечивающие морфологический анализ и нормализацию слов.

В качестве лексических единиц основных словарей используются ключевые слова, словосочетания и дескрипторы.

Дескриптор — понятие, обозначающее группу эквивалентных или близких по смыслу ключевых слов. Дескриптор - это имя класса синонимов. В качестве дескрипторов могут быть использованы код, слово или словосочетание.

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

Тезаурус в узком смысле представляет собой специальный словарь-справочник, в котором перечислены **ключевые слова** — дескрипторы определенной предметной области, указаны их синонимы, установлены способы устранения синонимии, омонимии, полисемии, определены родовидовые и ассоциативные связи дескрипторов.

Наиболее важными парадигматическими отношениями ИПТ являются:

- соподчинение;
- род-вид;
- часть—целое;
- причина-следствие;
- функциональное сходство.

Обобщенная структура ИПТ включает как минимум три составляющих: словарную часть, семантическую карту, руководство по использованию.

Словарная часть — алфавитный список дескрипторов с их словарными статьями.

Семантическая карта — система тематических классов дескрипторов, представленная в виде графической схемы или таблицы.

Руководство по использованию ИПТ содержит правила перевода ключевых слов и словосочетаний на ИПЯ, правила лексикографического контроля и редактирования поискового образа документа и поискового образа запроса, а также правила ведения ИПТ.

Отличием информационно-поисковых тезаурусов от информационно-поисковых каталогов на основе предметной иерархической рубрикации является то, что в тезаурусах, помимо классификационной схемы, присутствуют сами ключевые слова и дескрипторы, объединяемые под названием классов, рубрик и т. д. В каталогах же присутствуют только лишь обозначения (названия) классов.

Главная идея информационно-поисковых тезаурусов заключается в повышении эффективности индексирования документов в рамках дескриптивного подхода. Однако в процессе индексирования учитываются

семантические отношения между дескрипторами, что, в конечном счете, обеспечивает более адекватный содержанию документа поисковый образ и повышает эффективность поиска документов.

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:

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).

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.

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.

Итоги теоретического исследования позволили нам выявить достоинства и недостатки различных ДИПЯ, базирующихся на том или ином информационно-поисковым языке.

Так как каждая поисковая система предоставляет различные возможности поиска, из различных баз данных, поэтому информационный поиск на базе ИПС представляет собой достаточно сложный процесс познавательно-практической деятельности, требующий от поисковых субъектов априорной подготовки.

Анализ ИПЯ сети интернет поможет провести свой собственный выбор наиболее подходящего средства поиска, которое обеспечивало актуальность, быстроту и точность результатов.

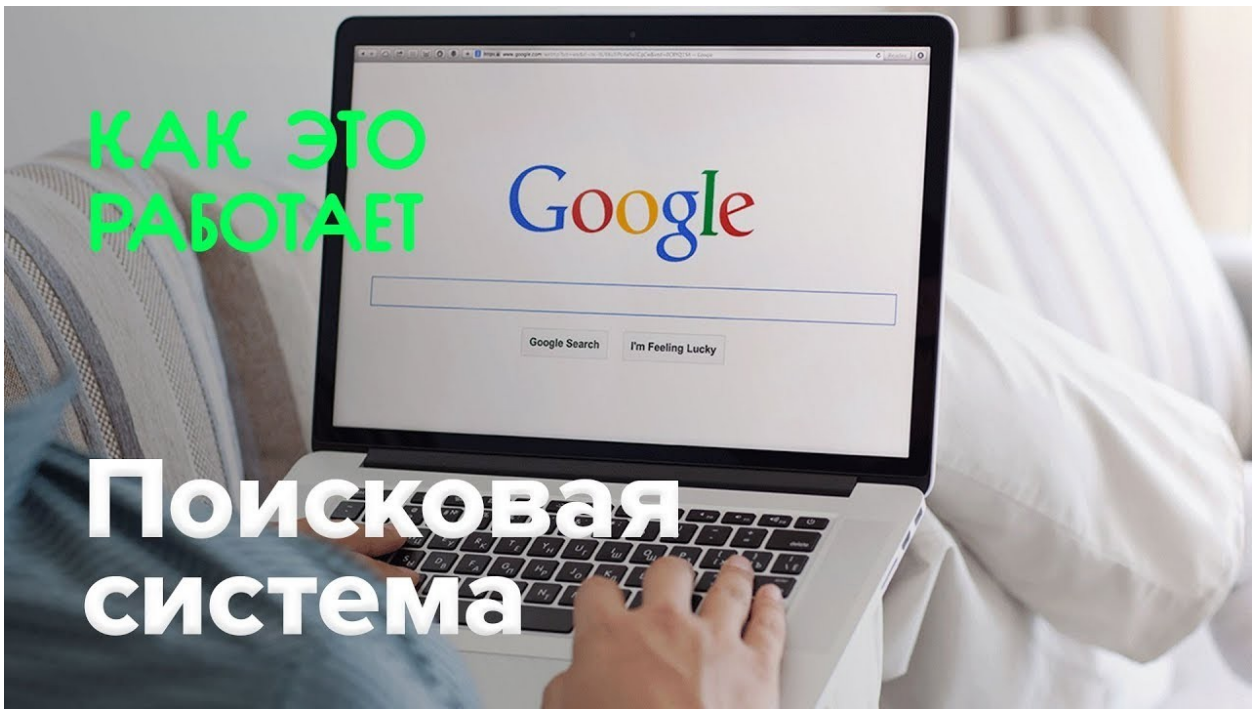
Библиография.

1. Гост 7.25-2001 СИБИД. Информационно-поисковые языки. Термины и определения - М.: Изд-во стандартов,2001.-38 с. Воробьёв Г.Г. Документ: информационный анализ. М., 1973. С.

2. Захаров В.П. Информационно-поисковые системы: Учебно-методическое пособие.- Спб.,2005 г,48 с. Современное делопроизводство. Кирсанова М.В. М: Инфра – М 2000.

3. Храмцов П.И. Информационно-поисковые языки, М.: «Гелиос», 2008


HTML5
PHP
</>
{}


КАК ЭТО РАБОТАЕТ
Google
Google Search
I'm Feeling Lucky
Поисковая система