Лекция 3

Тема: Проектирование документальных баз данных. Информационно-поисковые системы

Цель: Изучить основы проектирования документальных баз данных.

План:

1. Документальная система.
2. Общая функциональная структура документальных информационно-поисковых систем.
3. Информационно-поисковые языки. Обработка входящей текстовой информации.
4. Поиск текстовой информации. Модели поиска текстовой информации (булева модель, модель нечетких множеств, пространственно-векторная модель, вероятностные модели).

Литература:

1. Базы данных. Интеллектуальная обработка информации / В.В. Корнеев, А.Ф. Гареев, С.В. Васютин, В.В. Райх. М., 2000. С. 69-89.
2. Кузин А.В. Базы данных: Учеб. пособие для студ. высш. учеб. за-ведений/А. В. Кузин, С.В. Левенисова.-М.: ИЦ " Академия", 2008.-320 с.

1. Документальная система

В отличие от фактографических баз данных, ориентированных на полное и точное представление данных достаточно простой смысловой структуры, документальные БД ориентированы на частичное, приближенное представление данных, имеющих значительно более сложную смысловую структуру, представленных на входе в форме текста [1, 2].

Основной функцией любой ДИПС является информационное обеспечение потребителей на основе выдачи ответов на их запросы. Осуществление выдачи системой требуемых данных реализуется с помощь главной операции ДИПС - проведения информационного поиска [1].

Информационный поиск является процедурой отыскания документов, содержащих ответ на заданные потребителем вопросы.

Информационный поиск в системе проводится на основе поступившего от потребителя запроса на отыскание необходимой ему информации. Потребность человека в определенной информации в процессе его практической деятельности носит название *информационной потребности.* Под действием получаемой информации информационная потребность людей постоянно изменяется и трансформируется. Вследствие этого её невозможно однозначно выразить и описать. Однако информационная потребность может быть представлена в виде некоторой последовательности ее частных значений в фиксированные моменты времени. Такое частное значение информационной потребности потребителя в определенные моменты времени, выраженное на естественном языке (ЕЯ), и

представляет собой *информационный запрос,* с которым пользователь обращается к системе [1].

Следовательно, реакцию системы необходимо рассматривать не только по отношению к информационной потребности, но по отношению к информационному запросу. Для выражения данных отношений в теории ДИПС введены два фундаментальных понятия: пертинентность и релевантность.

Под *пертинентностью* понимается соответствие смыслового содержания документа информационной потребности потребителя. Документы, содержание которых удовлетворяет информационной потребности, называют пертинентными.

Релевантность представляет собой соответствие содержания документа информационному запросу в том виде, в каком он сформулирован, а документы, содержание которых отвечает запросу потребителя, носят название релевантных.

Автоматизация процесса информационного поиска потребовала формализации представления основного смыслового содержания информационного запроса и документов в виде соответственно *поискового предписания (ПП)* и *поисковых образов документов (ПОД)*. Для записи ПП и ПОД применяются специальные языки, называемые информационно-поисковыми (или просто информационными).

В процессе проведения информационного поиска в ДИПС определяется степень соответствия содержания документов и запроса пользователя путем сопоставления ПОД с ПП. А на основе такого сопоставления принимается решение о выдаче документа (он признается релевантным) или его невыдаче (он считается нерелевантным).

Решение о выдаче или невыдаче документа в ответ на запрос принимается на основе некоторого набора правил, по которому данной ДИПС определяется степень смысловой близости между ПОД и ПП. Такой набор правил получил название *критерия смыслового соответствия* (КСС).

2. Общая функциональная структура документальных информационно-поисковых систем

В состав типичной ДИПС входят, как правило, четыре основные подсистемы (рис. 1) [1]:

1. Подсистема ввода и регистрации.
2. Подсистема обработки.
3. Подсистема хранения.
4. Подсистема поиска.

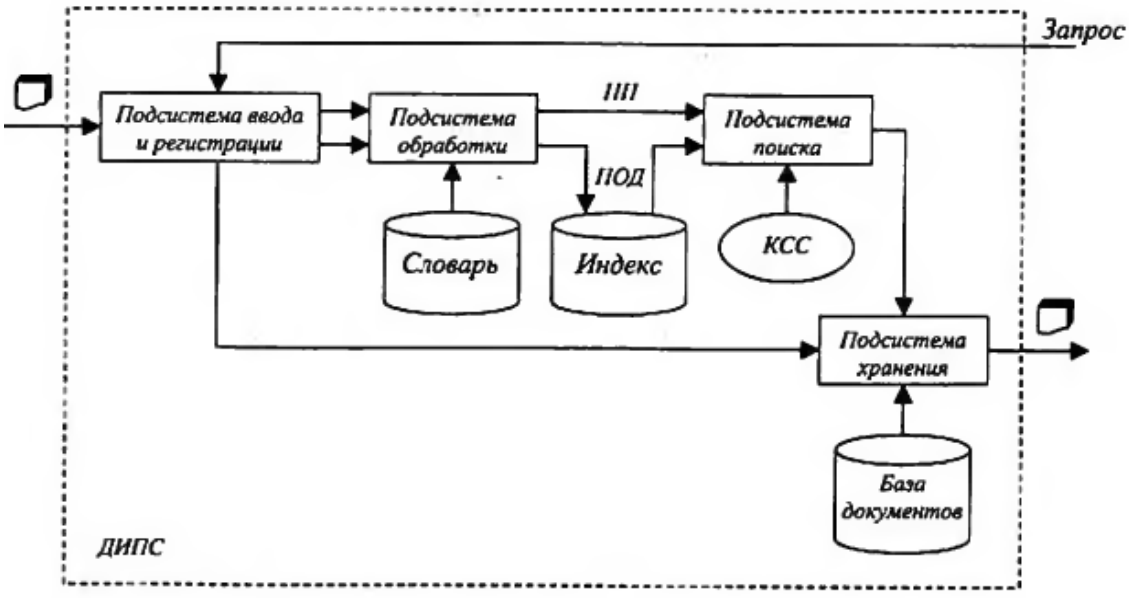

Рис. 1. *Общая функциональная структура ДИПС* [1]

Подсистема ввода и регистрации решает следующие основные задачи:

- создание электронных копий бумажных документов (например, сканирование с последующим распознаванием текста или ввод с клавиатуры);
- обеспечение подключения к каналам доставки электронных документов;
- распознавание, а при необходимости и преобразование формата электронных документов;
- присвоение электронным документам уникальных идентификаторов (регистрация), а также ведение таблицы синхронизации имен (при необходимости сохранения прежних имен).

Все поступающие документы без внесения в них каких-либо изменений направляются в ***подсистему хранения*** для сохранения в базе документов. База документов может представлять собой простую совокупность файлов, распределенную по каталогам жесткого диска.

Для хранения документов применяют средства сжатия и быстрого поиска информации. В этом случае подсистема хранения представляет собой совокупность стандартных или специализированных средств архивации, СУБД и т.п., обеспечивающих возможность доступа к данным по предъявляемому идентификатору.

Далее документы поступают на вход ***подсистемы обработки***, задачей которой является формирование для каждого документа ПОД, в который заносится информация, необходимая для последующего поиска документа.

ПОД сохраняются в индексе. Логически индекс представляет собой таблицу, строки которой соответствуют документам, а столбцы - информационным признакам, на основе

которых строится ПОД. В ячейках таблицы могут храниться либо 1, либо 0 – в зависимости от наличия или отсутствия данного признака в данном документе.

При поступлении на вход системы запроса пользователя он преобразуется в ПП и передается в ***подсистему поиска***, задачей которой является отыскание в индексе ПОД, удовлетворяющих ПП с точки зрения КСС. Идентификаторы релевантных документов подаются с выхода подсистемы поиска на вход подсистемы хранения, которая осуществляет выдачу пользователю самих релевантных документов.

3. Информационно-поисковые языки. Обработка входящей текстовой информации

Информационно-поисковым языком (ИПЯ) называется специализированный искусственный язык, предназначенный для описания основного смыслового содержания поступающих в систему сообщений, с целью обеспечения возможности последующего их поиска.

ИПЯ принято разбивать на два основных типа:

- классификационные языки,
- дескрипторные языки.

С помощью классификационных языков производится классификация сообщений, т.е. отнесение их к классам, обозначенным лексическими единицами (ЛЕ) ИПЯ. Частным случаем классификационного ИПЯ является рубрикатор, лексическими единицами которого являются названия тематических рубрик. В целом *под рубрикатором* некоторой предметной области понимается ориентированный граф, состоящий из независимых деревьев. Листья деревьев будем *называть рубриками* - объектами, инкапсулирующие знания о конкретных фрагментах данной предметной области. Все нелистовые вершины являются классификационными родово-видовымн обобщениями листовых вершин и используются лишь при ведении информационного поиска.

В дескрипторных ИПЯ ЛЕ заранее не связаны никакими текстуальными отношениями. Сложные синтаксические конструкции - предложения или фразы - создаются в этих языках путем объединения (координации) ЛЕ во время процедуры представления смыслового содержания документов системы. Готовых предложений или фраз в таких языках нет, поэтому отсутствуют ограничения на составление сложных понятий. Фактически из небольшого числа ЛЕ данные языки позволяют строить предложения, выражающие практически любой смысл. Такие ИПЯ носят также название посткоординируемых, поскольку координация между словами предложения возникает во время его записи. [1]

Обработка входящей текстовой информации

Т.к. документы, поступающие на вход ДИПС, записаны на ЕЯ, в ней обязательно должна проводиться операция перевода текстов входных документов с ЕЯ на ИПЯ. Тип используемого ИПЯ оказывает сильное влияние на суть процессов обработки информации в конкретных ДИПС. В случае применения ИПЯ дескрипторного типа такая операция перевода называется *индексированием,* при использовании рубрикатора *рубрицированием.*

На сегодняшний день среди дескрипторных ИПЯ наибольшее распространение в автоматизированных ДИПС получили языки без грамматики и без контроля по словарю. При их использовании говорят о *полнотекстовом индексировании.*

В операции перевода можно выделить два этапа:

1. Анализ смыслового содержания текста с целью выделения из него сведений об известных системе объектах, их свойствах, а также отношениях между ними.
2. Выражение этих сведений на ИПЯ, т.е. принятие решения о приписывании данному сообщению выражений на ИПЯ (о включении соответствующих выражений на ИПЯ в ПОД).

Лингвистический анализ текста

Лингвистический анализ текста может состоять из двух этапов:

1. морфологического анализа;
2. синтаксического анализа.

Цель *морфологического анализа* состоит в получении основ (под основой понимается словоформа с отсеченным окончанием) со значениями грамматических категорий (например, часть речи, род, число, падеж) для каждой из словоформ.

Задачей *синтаксического анализа* является осуществление грамматического разбора предложений на основе информации, заложенной в словаре. На этом этапе выделяется подлежащее, сказуемое, дополнение и т.п., между которыми указываются связи по управлению в виде дерева зависимостей.

Автоматическое индексирование

Автоматическое индексирование документов может основываться на простых, однословных или многословных составных терминах (фразах). Простые, однословные термины далеко не идеальны для индексирования, поскольку смысл слов вне контекста нередко бывает неоднозначным. Термины-фразы более осмысленны, обладают большей дискриминирующей мощью. Для генерации фраз может использоваться как синтаксический анализ, так и ряд эвристических алгоритмов. Ниже приведено описание одного из них.

4. Поиск текстовой информации. Модели поиска текстовой информации

Модель поиска текстовой информации характеризуется четырьмя параметрами [1]:

- представлением документов и запросов;
- критерием смыслового соответствия;
- методами ранжирования результатов запроса;
- механизмами обратной связи, обеспечивающими оценку релевантности пользователем.

Рассмотрим наиболее распространенные модели поиска с позиции первых трех параметров.

Булева модель представляет документы с помощью набора терминов, присутствующих в индексе, каждый из которых рассматривается как булева переменная. При наличии термина в документе соответствующая переменная принимает значение True. Присваивание терминам весовых коэффициентов не допускается. Запросы формулируются как произвольные булевы выражения, связывающие термины с помощью стандартных логических операций: AND, OR или NOT. Мерой соответствия запроса документу служит значение статуса выборки (RSV, retrieval status value). В булевой модели RSV равно либо 1, если для данного документа вычисление выражения запроса дает True, либо 0 в противном случае. Все документы с RSV = 1 считаются релевантными запросу.

Такая модель проста в реализации и применяется во многих коммерческих системах. Она позволяет пользователям вводить в свои запросы произвольные сложные выражения. Однако эффективность поиска обычно невысока. К тому же, ранжировать результаты невозможно, так как все найденные документы имеют одинаковые RSV, а терминам нельзя присвоить весовые коэффициенты. Нередко результаты выглядят противоестественно. Например, если пользователь указал в запросе десять терминов, связанных логической операцией AND, документ, содержащий девять таких терминов, в выборку не попадет. Для повышения эффективности поиска в ИПС часто применяется обратная связь с пользователем. Как правило, система просит пользователя указать релевантность или нерелевантность нескольких документов, включенных в начало списка вывода. Поскольку результаты не ранжируются, выбор документов для подобной экспертной оценки релевантности затруднен.

Модель нечетких множеств основывается на теории нечетких множеств, допускающей (в отличие от обычной теории множеств) частичную принадлежность элемента тому или иному множеству. Здесь логические операции переопределены таким образом, чтобы учесть возможность неполной принадлежности множеству, а обработка

запросов пользователя выполняется аналогично булевой модели. Тем не менее, ИПС на основе подобной модели оказывается практически столь же не способной классифицировать полученные результаты, что и системы, базирующиеся на булевой модели.

Строгая булева модель и модель, использующая методы теории нечетких множеств, требуют меньших объемов вычислений (при индексировании и оценке соответствия документов запросу), чем другие модели. Они менее сложны алгоритмически и предъявляют не очень жесткие требования к другим ресурсам, таким как дисковое пространство для хранения представлений документов.

Пространственно-векторная модель основана на предположении, что совокупность документов можно представить набором векторов в пространстве, определяемом базисом, из n нормализованных векторов терминов. Значение первого компонента вектора представляющего документ отражает вес термина в нем. Запрос пользователя также представляется n-мерным вектором. Показатель RSV, определяющий соответствие документа запросу, задастся скалярным произведением векторов запроса и документа. Чем больше RSV, тем выше релевантность документа запросу.

Достоинство подобной модели в ее простоте. Она позволяет легко реализовать обратную связь для оценки релевантности пользователем. В то же время приходится жертвовать выразительностью спецификации запроса, присущей булевой модели.

Вероятностные модели. В пространственно-векторной модели подразумевается, что векторы терминов, ортогональны и существующие взаимосвязи между терминами не должны приниматься во внимание. Кроме того, в такой модели не специфицируется степень соответствия "запрос - документ" и она оценивается достаточно произвольно. Вероятностная модель учитывает все взаимозависимости и связи терминов, а также определяет такие основные параметры, как веса терминов запросов и форма соответствия «запрос-документ».

Данная модель базируется на двух главных параметрах: Pr (rel) и Pn (nonrel), т.е. вероятности релевантности и нерелевантности документа запросу пользователя, которые вычисляются на основе вероятностных весовых коэффициентов терминов и фактического присутствия терминов в документе.

Вопросы для самопроверки:

1. Классификация БД по типу хранимой информации.
2. Каково назначение документальных систем?
3. Что такое информационный поиск? информационная потребность? информационный запрос? пертенентность? релевантность?
4. Что такое поисковое предписание и поисковый образ документа?

5. Какова структура документальной информационно-поисковой системы (ИПС)?
6. Каково назначение подсистемы ввода и регистрации? подсистемы обработки? подсистемы хранения? подсистемы поиска?
7. Что такое информационно-поисковый язык?
8. Как обрабатывается входящая текстовая информация в ИПС?
9. Какими параметрами характеризуется модель поиска текстовой информации?
10. Как работа ИПС на основе булевой модели?
11. Как работа ИПС на основе модели нечетких множеств?
12. Как работа ИПС на основе пространственно-векторной модели?
13. Как работа ИПС на основе вероятностной модели?
14. Какие методы введения обратной связи с пользователем в ИПС вы знаете?