Информационные системы

Тема 6: «Документальные информационные системы»

Колмыкова Оксана Владимировна

ВГУЭС, г.Владивосток, ул.Гоголя 41

каф. ИИКГ, ауд. 1448

Документальная информационная система

единое хранилище документов с инструментарием поиска и выдачи необходимых пользователю документов.

Поисковый характер документальных информационных систем исторически определил еще одно их название — *информационно-поисковые системы* (ИПС), хотя этот термин не совсем полно отражает специфику документальных информационных систем.

Единичным элементом данных

в документальных информационных системах является *неструктурированный на более мелкие элементы* ***документ.***

В качестве неструктурированных документов в подавляющем большинстве случаев выступают, прежде всего, ***текстовые документы,*** представленные в виде текстовых файлов, хотя к классу неструктурированных документированных данных могут также относиться звуковые и графические файлы.

Основная задача документальных информационных систем

хранение, накопление и предоставление пользователю документов, содержание, тематика, реквизиты которых соответствуют его информационным потребностям.

Соответствие найденных документов информационным потребностям пользователя называется ***пертинентностью.***

В силу теоретических и практических сложностей с формализацией смыслового содержания документов пертинентность относится скорее к качественным понятиям, хотя, как будет рассмотрено ниже, может выражаться определенными количественными показателями.

В зависимости от особенностей реализации хранилища документов и механизмов поиска документальные ИПС можно классифицировать на две группы:

- системы на основе индексирования;
- семантически-навигационные системы.

Семантически-навигационные системы

Документы, помещаемые в хранилище (в базу) документов, оснащаются специальными *навигационными конструкциями*, соответствующими *смысловым связям* между различными документами или отдельными фрагментами одного документа.

Такие конструкции реализуют некоторую *семантическую* (смысловую) *сеть* в базе документов.

Способ и механизм выражения информационных потребностей в подобных системах заключаются в явной навигации пользователя по смысловым отсылкам между документами.

В настоящее время такой подход реализуется в ***гипертекстовых информационно-поисковых системах.***

Системы 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.***

Системы на основе индексирования

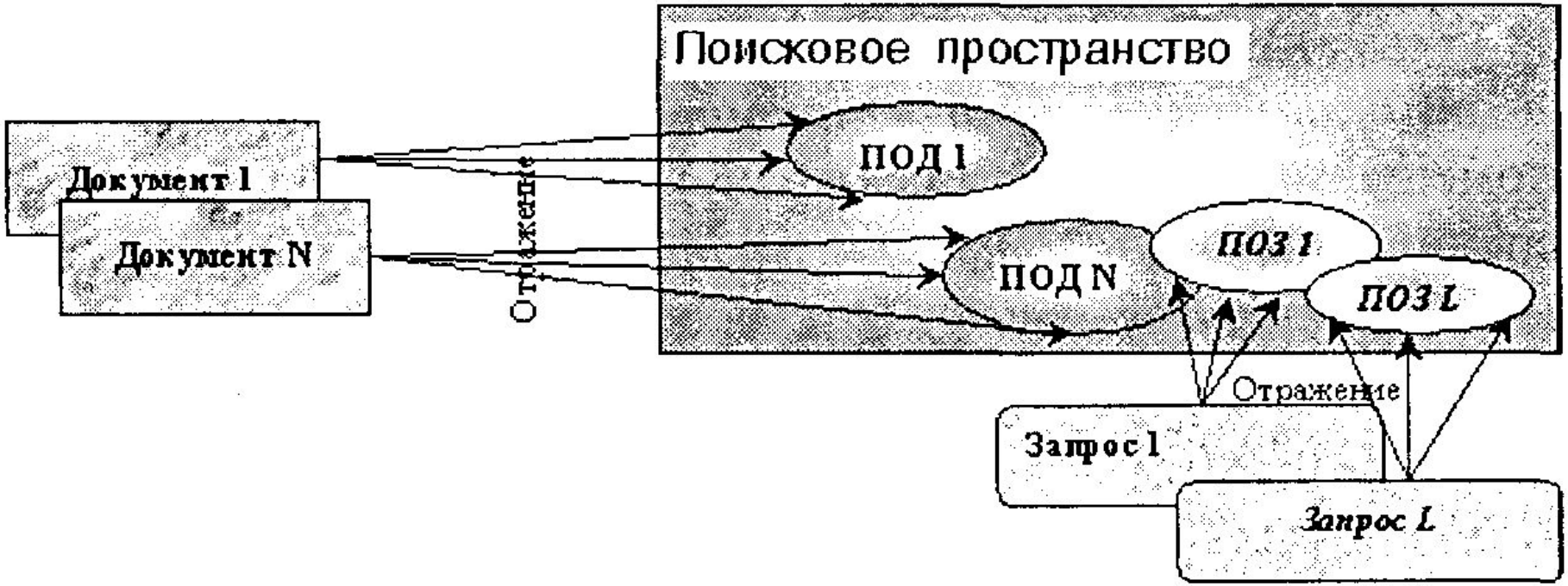
Процесс отображения документа в поисковое пространство называется ***индексированием*** и заключается в присвоении каждому документу некоторого индекса-координаты в поисковом пространстве. Формализованное представление (описание) индекса документа называется ***поисковым образом документа*** **(ПОД).**

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 или близки поисковым образам запроса пользователя.

Соответствие найденных документов запросу пользователя называется ***релевантностью.***

Общий принцип устройства и функционирования документальных ИПС на основе индексирования

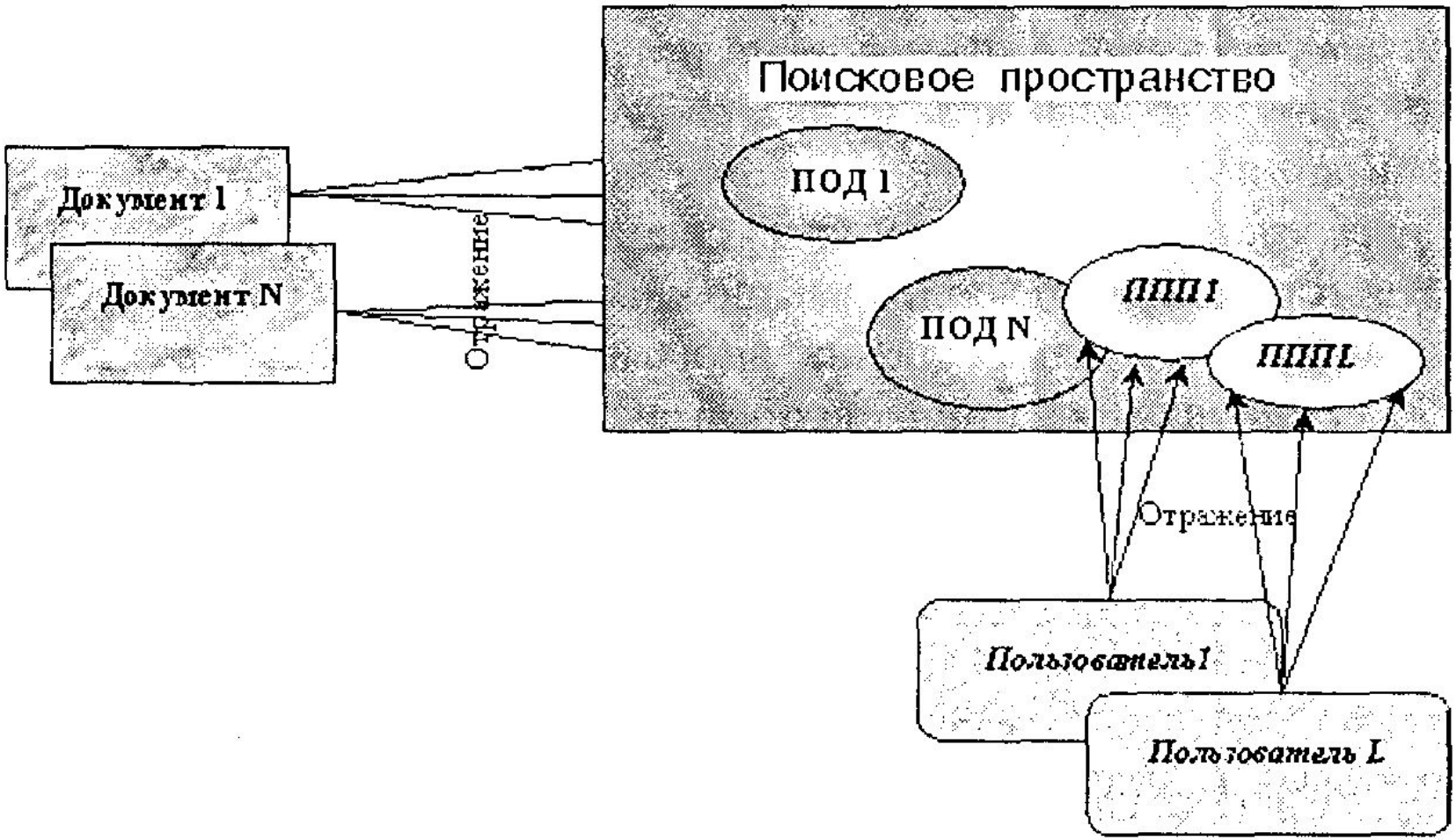


Принцип решения задач **информационного оповещения** в документальных ИПС на основе индексирования

аналогичен принципу решения задач поиска документов по запросам и основан на *отображении в поисковое пространство информационных потребностей пользователя в виде* так называемых ***поисковых профилей пользователей*** (ППП).

Информационно поисковая система по мере поступления и индексирования новых документов сравнивает их образы с поисковыми профилями пользователей и принимает решение о соответствующем оповещении.

Принцип решения задач информационного оповещения в документальных ИПС



Поисковое пространство

отображает поисковые образы документов и реализующие механизмы информационного поиска документов, строится на основе *языков документальных баз данных,* называемых **информационно-поисковыми языками (ИПЯ).**

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

Информационно-поисковый язык можно разделить на составляющие:

- структурная;
- манипуляционная.

Структурная составляющая ИПЯ

Структурная составляющая ИПЯ документальных ИПС на основе индексирования реализуется ***индексными указателями*** в форме:

1) информационно-поисковых каталогов,
2) тезаурусов,
3) генеральных указателей.

1) Информационно-поисковые каталоги

являются традиционными технологиями организации информационного поиска в документальных фондах библиотек, архивов и представляют собой *классификационную систему знаний по определенной предметной области.*

Смысловое содержание документа в информационно-поисковых каталогах *отображается* тем или иным *классом каталога,* а *индексирование* документов заключается в *присвоении* каждому документу специального *кода (индекса)* соответствующего по содержанию класса (классов) каталога и создания на этой основе специального индексного указателя.

2) Тезаурус

представляет собой специальным образом *организованную совокупность основных лексических единиц (понятий) предметной области (словарь терминов) и описание парадигматических отношений между ними.*

Парадигматические отношения выражаются *семантическими отношениями между элементами словаря, не зависящими от любого контекста.*

Так же, как и в информационно-поисковых каталогах, в системах на основе тезаурусов в информационно-поисковое пространство отображается не весь текст документа, а только лишь выраженное средствами тезауруса смысловое содержание документа.

3) Генеральный указатель (глобальный словарь-индекс)

в общем виде представляет собой *перечисление всех слов (словоформ), имеющихся в документах хранилища, с указанием координатного местонахождения каждого слова*.

Индексирование нового документа в таких системах производится через дополнение *координатных отсылок тех словоформ генерального указателя, которые присутствуют в новом документе.*

Так как поисковое пространство в таких системах *отражает полностью весь текст документа* (все слова документа), а не только его смысловое содержание, то такие системы получили название ***полнотекстовых ИПС*.**

Структурная составляющая ИПЯ семантически-навигационных систем

реализуется в виде техники смысловых отсылок в текстах документов и специальном навигационном интерфейсе по ним и в настоящее время представлена ***гипертекстовыми технологиями.***

Манипуляционная составляющая ИПЯ

Поисковая (манипуляционная) составляющая ИПЯ

реализуется

дескрипторными

и

семантическими

языками запросов.

Дескрипторные языки запросов

Документы и запросы представляются *наборами некоторых лексических единиц — дескрипторов, не имеющих между собой связей,* или, как еще говорят, *не имеющих грамматики.*

Таким образом, каждый *документ* или запрос *представлен* некоторым *набором дескрипторов*.

Поиск осуществляется через поиск документов с *подходящим набором дескрипторов*.

В качестве элементов-дескрипторов выступают либо элементы ***словаря ключевых терминов***, либо элементы ***генерального указателя*** (глобального словаря всех словоформ).

В силу отсутствия связей между дескрипторами, набор которых для конкретного документа и конкретного запроса выражает, соответственно, поисковый образ документа (ПОД) или поисковый образ запроса (ПОЗ), такие языки применяются, прежде всего, в полнотекстовых системах.

Семантические языки запросов

содержат грамматические и семантические конструкции для выражения (описания) смыслового содержания документов и запросов.

Все многообразие семантических языков подразделяется на две большие группы:

- предикатные языки;
- реляционные языки.

Классификационные системы поиска документов

Особенностью систем перечислительной классификации является **возможность индексирования документов любым количеством предметов (рубрик),** отражающих содержание документа.

Для осуществления поиска необходимых документов по классификатору (каталогу) определяются коды интересующих абонента предметов (рубрик) и далее отбираются из хранилища те документы, которые проиндексированы соответствующими кодами.

Для удобства поиска и отбора по каждому документу формируется специальная карточка, на которую наносится информация о кодах предметных рубрик документа, а также, о его физическом местонахождении, и реферат, который уже на естественном языке в сжатом виде отражает содержание документа.

Поиск и отбор документов непосредственно осуществляется по отбору карточек с необходимыми индексными кодами для последующего извлечения из хранилища собственно самих документов

Основные показатели эффективности функционирования информационно-поисковых систем

- ***Полнота информационного поиска R*** определяется отношением числа найденных пертинентных документов *A* к общему числу пертинентных документов C, имеющихся в системе или в исследуемой совокупности документов
- ***Точность информационного поиска P*** определяется отношением числа найденных пертинентных документов *A* к общему числу документов *L,* выданных на запрос пользователя
- ***Коэффициент информационного шума к,*** соответственно, определяется отношением числа нерелевантных документов *(L-A),* выданных в ответе пользователю к общему числу документов *L,* выданных на запрос пользователя

СПАСИБО ЗА ВНИМАНИЕ!!!

Контрольные вопросы № 6

1. Дайте определение документальной информационной системы.
2. Перечислите классификационные системы поиска документов.
3. В чем заключается основная задача документальных информационных систем?
4. Дайте определение дескриптора.
5. Что является единичным элементом данных в документальных информационных системах?
6. Дайте определение поискового образа документа.
7. Какой механизм поиска документов реализуется в гипертекстовых информационно-поисковых системах?
8. Что понимают под пертинентностью?
9. Как определяется точность информационного поиска?