# ИНФОРМАЦИОННЫЕ СИСТЕМЫ

*Лекция 2*

*Документальные и фактографические ИС*

# Содержание

- АИС
- Документальные и фактографические ИС
- Документальные ИС
- Пертинентность и релевантность
- Функциональная структура ДИПС
- Информационно-поисковые языки
- Оценка качества ДИПС

# АИС

В 60-х годах была осознана необходимость применения средств компьютерной обработки хранимой информации там, где были накоплены значительные объемы полезных данных – в военной промышленности, в бизнесе. Появились **автоматизированные информационные системы** (АИС) – программно-аппаратные комплексы, предназначенные для хранения, обработки информации и обеспечения ею пользователей.

# АИС

Первые АИС работали преимущественно с информацией *фактического* характера, например, характеристиками объектов и их связей. По мере «интеллектуализации» АИС появилась возможность обрабатывать текстовые документы на естественном языке, изображения и другие виды и форматы представления данных.

Принципы хранения данных в системах обработки фактической и документальной (текстовой) информацией схожи, но алгоритмы обработки заметно отличаются.

# Документальные и фактографические ИС

Поэтому в зависимости от характера информационных ресурсов, которыми оперируют такие системы, принято различать два крупных класса – **документальные и фактографические**.

**Документальные системы** служат для работы с документами на естественном языке – монографиями, публикациями в периодике, сообщениями пресс-агентств, текстами законодательных актов.

# Документальные и фактографические ИС

**Фактографические системы** оперируют фактическими сведениями, представленными в виде специальным образом организованных совокупностей формализованных записей данных. Центральное функциональное звено фактографических информационных систем – системы управления базами данных (СУБД).

# Документальные ИС

Классические модели и методы в теории ИС изначально ориентировались на организацию хранения и обработки детально структурированных данных.

Однако, на практике оказалось, что информация чаще представлена в виде простых текстовых документов.

Итак, ДИС – это системы, ориентированные на работу с текстовыми документами, с данными, имеющими приближенное представление, сложную структуру.

# Документальные ИС

Наиболее распространенный тип документальных систем – информационно-поисковые системы (ДИПС), предназначенные для накопления и поиска по различным критериям документов на естественном языке.

# Документальные ИС

В отличие от ФИПС, которые в ответ на запрос потребителя осуществляют выдачу конкретных сведений (фактов), ДИПС в результате поиска предоставляет потребителю совокупность документов, смысловое содержание которых соответствует запросу.

# Документальные ИС

Потребность человека в определенной информации в процессе его практической деятельности носит название **информационной потребности**.

Частное значение информационной потребности в определенные моменты времени, выраженное на ЕЯ, представляет собой **информационный запрос**, с которым пользователь обращается к системе.

# Документальные ИС

В теории ДИПС введены два фундаментальных понятия: пертинентность и релевантность.

Документы, содержание которых удовлетворяет информационной потребности, называют **пертинентными** *(от англ. pertinence – уместность, связь, отношение)*.

**Релевантность** *(от англ. relevance – уместность)* представляет собой соответствие содержания документа информационному запросу в том виде, в каком он сформулирован.

# Документальные ИС

Автоматизация процесса информационного поиска потребовала формализации представления основного смыслового содержания информационного запроса и документов в виде соответственно поискового предписания (ПП) и поисковых образцов документов (ПОД).

Для записи ПП и ПОД применяются специальные информационно-поисковые языки.

# Документальные ИС

Решение о выдаче или невыдаче документа в ответ на запрос принимается на основе некоторого набора правил, по которому данной ДИСП определяется степень смысловой близости между ПОД и ПП.

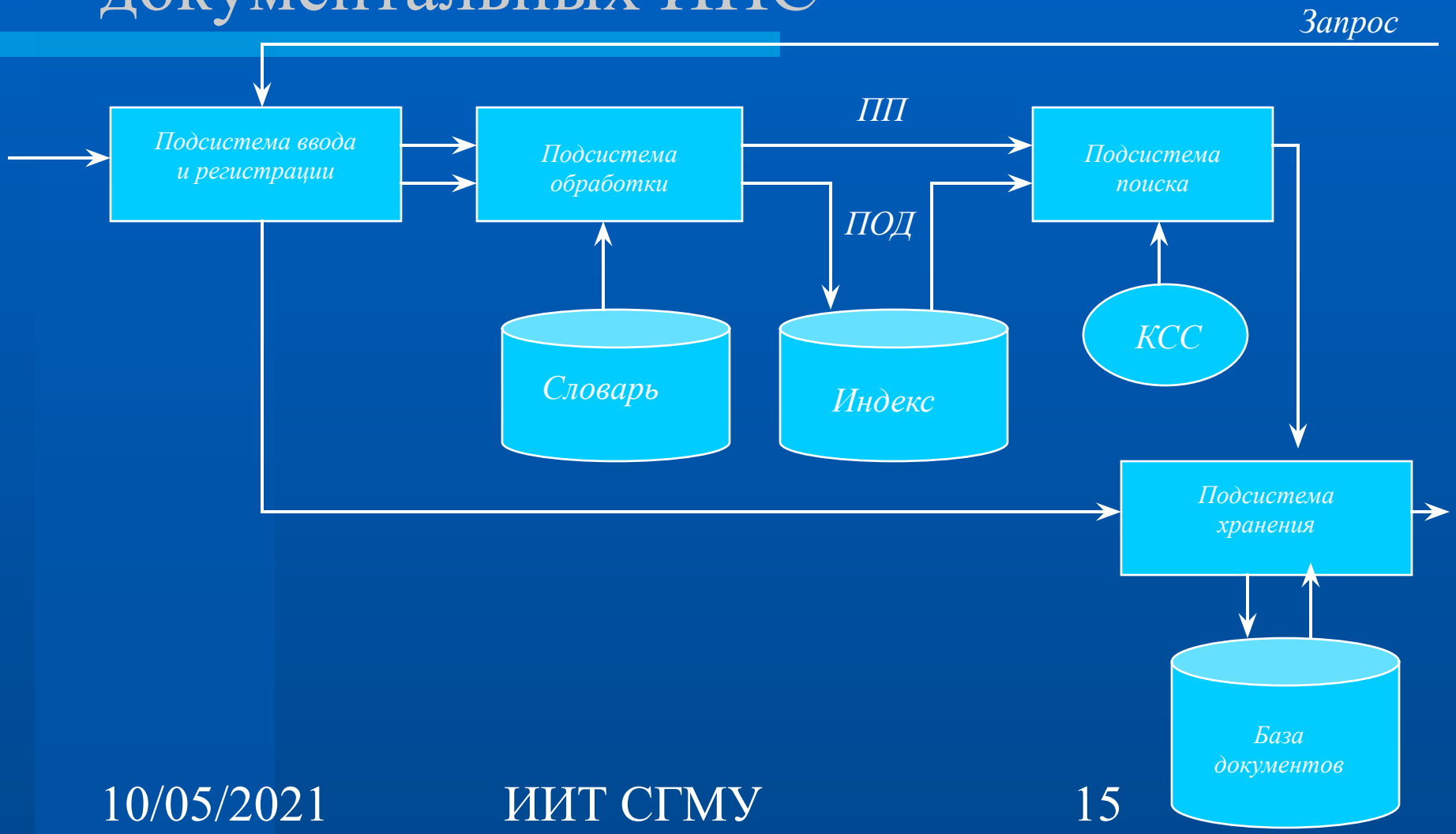
Такой набор правил получил название **критерия смыслового соответствия** (КСС).

# Общая функциональная структура документальных ИПС

В состав типичной ДИПС входят четыре основные подсистемы:

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

# Общая функциональная структура документальных ИПС



# Общая функциональная структура документальных ИПС

Задачи подсистемы ввода и регистрации:

- Создание электронных копий бумажных документов (сканирование, распознавание, ввод с клавиатуры);
- Обеспечение подключения к каналам доставки электронных документов;
- Присвоение эл.документам уникальных идентификаторов, ведение таблицы синхронизации имен;
- Преобразование формата эл.документов.

# Общая функциональная структура документальных ИПС

Для хранения документов применяют средства сжатия и быстрого поиска информации.

Система хранения:

- Средства архивации
- СУБД для доступа к данным по идентификатору.

Подсистема обработки формирует для каждого документа ПОД.

# Общая функциональная структура документальных ИПС

ПОД сохраняются в индексе. Логически индекс – таблица, строки которой соответствуют документам, а столбцы информационным признакам.

В ячейках таблицы могут хранится либо 1, либо 0 – в зависимости наличия или отсутствия данного признака в данном документе.

Такая таблица сильно разрежена, на практике хранят свертку таблицы по строкам и столбцам. Такую форму хранения называют прямой или инверсной.

# Общая функциональная структура документальных ИПС

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.

# Информационно-поисковые языки

Недостатки естественного языка (с точки зрения машинной технологии):

- Многообразие средств передачи смысла;
- Семантическая неоднозначность;
- Синонимия;
- Многозначность (полисемия – команда, омонимия - лук);
- Эллипсность (пропуски подразумеваемых слов).

# Информационно-поисковые языки

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

ИПЯ создается на базе ЕЯ, однако отличается от него компактностью, наличием четких грамматических правил и отсутствием семантической неоднозначности.

# Информационно-поисковые языки

ИПЯ принято разбивать на два основных типа:

- Классификационные языки
- Дескрипторные языки

Разница между данными типами – в процедуре построения предложений.

С помощью языков первого типа производится классификация сообщений.

# Информационно-поисковые языки

Например, частным случаем классификационного ИПЯ является рубрикатор.

Рубрикатор формируется группой экспертов, на основании их знаний о предметной области с учетом информационных потребностей пользователей.

Лексическими единицами являются названия тематических рубрик.

# Информационно-поисковые языки

В целом под рубрикатором некоторой предметной области понимается ориентированный граф, состоящий из независимых деревьев.

Листья деревьев будем называть рубриками – объектами, инкапсулирующими знания о конкретных фрагментах данной предметной области.

Все нелистовые вершины являются классификационными родово-видовыми обобщениями листовых вершин и используются лишь при ведении информационного поиска.

# Информационно-поисковые языки

Другой тип языков составляют дескрипторные ИПЯ, в которых ЛЕ заранее не связаны никакими текстуальными отношениями.

ДИПЯ различают с грамматикой и без грамматики. В первом случае имеет смысл порядок формирования синтаксических конструкций:

Иванов владеет автомобилем ---- владеть Иванов автомобиль

# Оценка качества ДИПС

В ПОД и ПП отражается лишь основное смысловое содержание поступающих сообщений в сокращенном виде. Поэтому метод поиска, основанный на сопоставлении ПП с ПОД, не в состоянии полностью обеспечить отыскания всех документов, отвечающих информационному запросу.

# Оценка качества ДИПС

Т.о., любой ДИСП присущи следующие ошибки:

- Ошибки 1-го рода (или пропуск цели): невыдача потребителю фактически релевантных его запросу документов;
- Ошибки 2-го рода (или ложная тревога, шум): выдача потребителю нерелевантных документов, которые не отвечают поставленному запросу.

# Оценка качества ДИПС

Разбиение массива документов:

| | Выданные | Невыданные |
|---|---|---|
| Релевантные | A | C |
| Нерелевантные | B | D |

Введем следующие обозначения:

a – кол-во выданных релевантных документов

b – кол-во выданных нерелевантных документов

c – кол-во невыданных релевантных документов

d – кол-во невыданных нерелевантных документов

# Оценка качества ДИПС

Существуют следующие показатели эффективности ДИПС:

1) Коэффициент полноты p, характеризующих долю выданных релевантных документов во всем массиве релевантных документов:

$$p = \frac{a}{a + c}$$

# Оценка качества ДИПС

2) Коэффициент точности n, характеризующих долю выданных релевантных документов во всем массиве выданных документов:

$$n = \frac{a}{a+b}$$

3) Коэффициент шума e, характеризующих долю выданных нерелевантных документов во всем массиве выданных документов:

$$e = \frac{b}{a+b} = 1 - n$$

# Оценка качества ДИПС

4) Коэффициент осадка q, характеризующих долю выданных нерелевантных документов во всем массиве нерелевантных документов:

$$q = \frac{b}{b+d}$$

5) Коэффициент специфичности k, характеризующих долю невыданных нерелевантных документов во всем массиве нерелевантных документов:

$$k = \frac{d}{b+d} = 1 - q$$

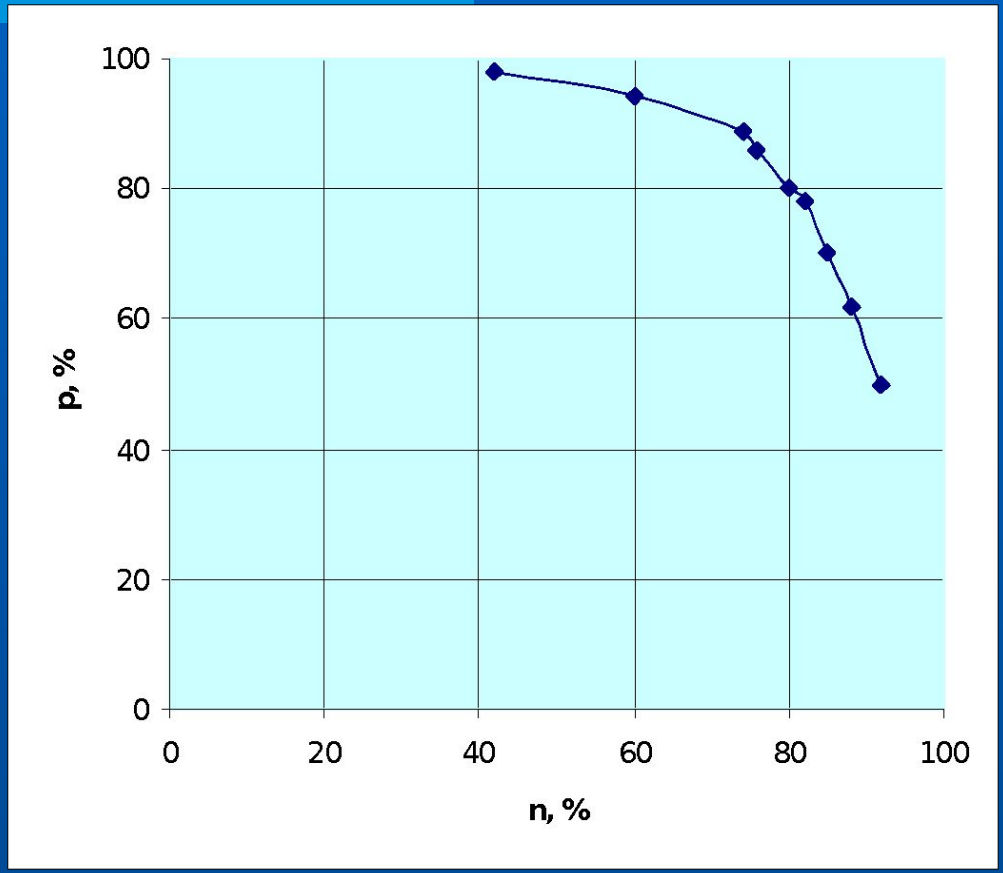
# Оценка качества ДИПС

Наиболее часто используются показатели полноты и точности.

Для удобства перечисленные показатели измеряют в %, у идеальной ДИСП полнота и точность 100%.

Однако такое качество поиска невозможно, поэтому на фиксированном уровне мощности поискового средства попытки улучшить один параметр приводят к ухудшению другого.

# Пример зависимости между p и n



# Оценка качества ДИПС

Другие показатели эффективности ДИПС:

- Быстродействие
- Пропускная способность
- Производительность (кол-во пользователей и частота их обращения)
- Надежность работы (оценивается вероятностью того, что система будет выполнять свои функции при заданных условиях в течение требуемого времени)
- Тип запросов, обслуживаемых системой

# Вопросы?