ДОКУМЕНТАЛЬНЫЕ СИСТЕМЫ

Подготовили студенты гр.9ИС-241

Березиков Артём Волков Илья

ОБЩЕЕ

На практике информация чаше всего представляется в виде текстовых документов, а не в виде структурированных данных. Документальные базы данных выделяются в один тип, который называется информационно-поисковая система (ИПС или ДИПС).

Документы ориентированы на приближенное представление данных. Главное назначение ДИПС — обработка запросов. Главная процедура — поиск (отыскание документа, содержащего ответ на запрос). При этом, в результате получается несколько документов. Запросы, как правило, формулируются на естественном языке (информационные запросы). Неправильно сформулированный запрос может не отражать информационные потребности пользователя.

Понятие пертинентность отражает смысловое соответствие документа информационным потребностям пользователя.

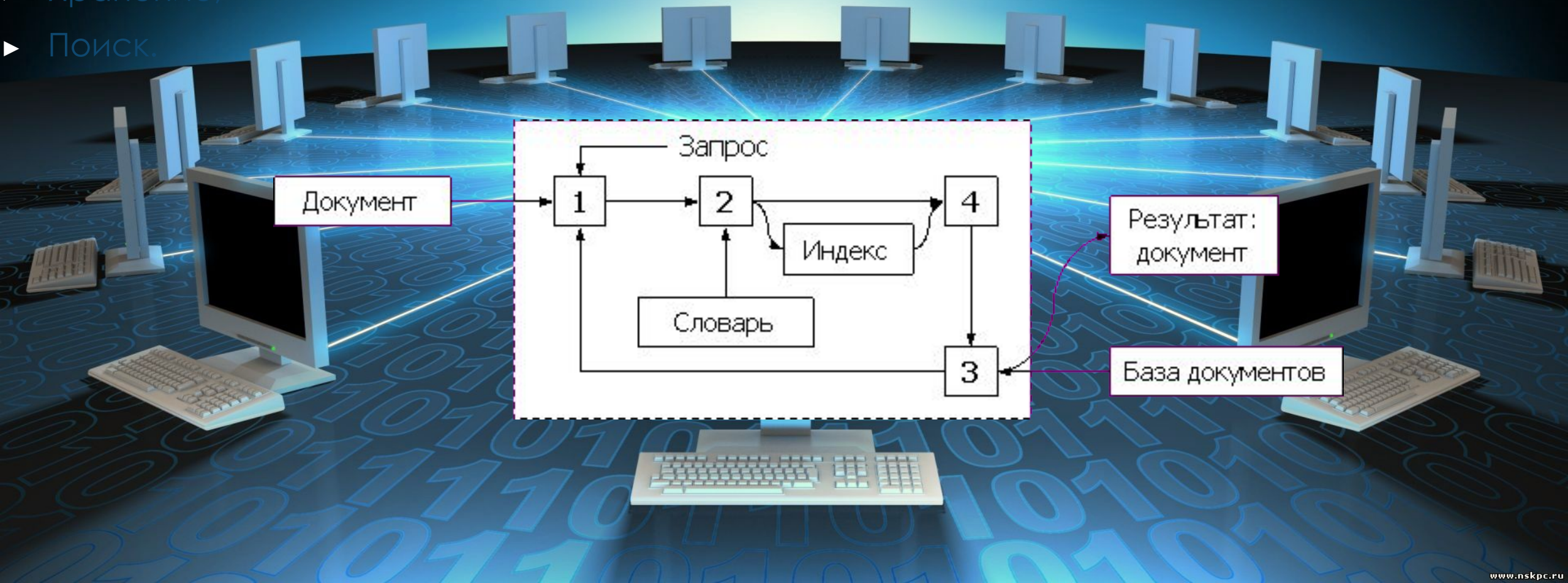
Релевантность — соответствие содержания документа информационному запросу в том виде, как он сформулирован.

Для автоматического поиска запросы представляются в виде информационного предписания — поискового предписания (ПП), а документы в виде поискового образа документа(ПОД).

Для записи ПП и ПОД применяется информационный поисковый язык. При поиске определяется соответствие ПП и ПОД, на основе которого принимается решение о выдаче документа, т.е. признания его релевантным. При этом набор правил, по которому принимается решение, называется критерием смыслового соответствия (КСС).

Критерий может задаваться явно или неявно. КСС строится на основе формальной релевантности. Фактическую релевантность и пертинентность документов определяет пользователь.

- ***Структура ДИПС***
- В ДИПС входят 4 подсистемы:
- Ввод и регистрация;
- Обработка;
- Хранение;
- Поиск.



Подсистема ввода решает следующие вопросы:

•создание электронных копий (сканирование, распознавание, ввод с клавиатуры);

•подключение к каналам доставки электронных документов (электронная почта и т.д.);

•преобразование форматов электронных документов;

•присваивание электронным документам уникальных идентификаторов;

•синхронизация имен.

•Подсистема обработки формирует для каждого документа поисковый образ, который необходим для дальнейшего поиска. Образец хранится в индексе (индекс-таблице). Индекс- это таблица, в строках которой стоят IDдокумента, а в столбцах — информационные признаки, на основе которых строится данный образ документа. Данные, как правило, бинарные. Поскольку таблицы сильно разряжены, то для их хранения обычно используют свертку. Запрос пользователя в системе преобразуется в поисковое предписание (ПП) и передается в систему поиска.

Формальное представление семантики документов

- Естественный язык не может быть использован в качестве представления информации из-за следующих недостатков:
- Многообразие передачи смысла, обеспеченное лексикой языка, контекстом, текстуальными отношения между словами, ссылками на раннее упомянутые слова.
- Семантическая неоднозначность слов.
- Синонимия, антонимия.
- Многозначность (полисемия) совпадение написания похожих слов.
- Эллипсность — пропуск подразумеваемых слов.

Невозможность использования естественных языков для поиска информации привело к созданию информационно-поисковых языков (ИПЯ). Эти языки применяются для смыслового описания текста, с целью последующего поиска. Они строятся на базе естественных языков, но отличаются четкими грамматическими правилами и отсутствием неоднозначностей. Все языки в ИПС делятся на два класса

- классификационные
- дескрипторные (словарные)

Обработка входящей информации в ДИПС

- Так как документы поступают в систему в текстовом виде, то они должны быть преобразованы в ИПЯ. В случае применения классификационных языков применяется рубрицирование, в случае дескрипторных языков — индексирование. При этом в случае применения дескриптивных языков без грамматики и без контроля по словарю говорят ополнотекстовом индексировании.
- Подготовка текста проходит в два этапа:
- анализ семантики системы, анализ объектов и связей;
- выражение этих связей на ИПЯ, приписывание объектами соответствующих выражений.

Поиск текстовой информации

Модель поиска характеризуется следующими параметрами:

1. Представление документов и запросов;
2. Критерий смыслового соответствия;
3. Методы ранжирования результатов запросов;
4. Механизм обратной связи для оценки релевантности документов.

Для представления документов и запросов применяется сразу несколько моделей.

Модели представления документов и запросов

Булева модель

В этом случае документ представляется с помощью набора терминов, присутствующих в индексе. Каждый термин представлен как булева переменная:

	Термин 1	Термин 2	Термин 3	
Документ 1	0	1	0	...
Документ 2	1	0	1	
Документ 3	1	1	0	
...				

Оценка качества ДИПС

В любой ДИПС присутствуют два типа ошибок:

1. Пропуск цели, т.е. невыдача релевантных документов;
2. Шум — выдача нерелевантных документов.

Весь массив документов можно разбить на 4 группы:

Документы	Выданные	Не выданные
Релевантные	A(a)	C(c)
Нерелевантные	B(b)	D(d)