ДОКУМЕНТАЛЬНЫЕ СИСТЕМЫ

1. Моделирование реальности в системах текстового поиска

От ИПС к системам текстового поиска

- *Информационно-поисковые системы*: поиск информации с помощью компьютеров

2 категории ИПС

- ***Фактографические:*** оперировали фактами, представленными в виде сущностей реального мира и их свойств.
- ***Документальные:*** предназначены для хранения и поиска документов, содержащих тексты на естественных языках.

Дескрипторные ИПС

- содержание каждого текстового документа и пользовательских поисковых запросов описывается наборами слов или словосочетаний, называемых *дескрипторами*.
- область применения: библиографический поиск.

Полнотекстовые документальные ИПС

- ***Полнотекстовыми*** называют системы, которые хранят и обрабатывают не описания документов, как это делается, например, в библиографических системах, а полные их тексты.
- ***Методы***: лингвистические, статистические.

- ***контекстный поиск -*** поиск документов, тексты которых содержат вхождение заданного в пользовательском запросе контекста.
- ***поиск по булевским критериям***.

Мультимедийные ИПС (системы текстового поиска)

- содержание их объектов поиска - составляет сочетание информационных ресурсов, представленных в различных средах - текстовых элементов, статических изображений, аудиоданных (музыкальные произведения, текст, произнесенный голосом и т.п.), мультфильмов, видео клипов и т.п.

- Охватывает большой спектр проблем - от теории информационного поиска до методов удовлетворения потребностей пользователей в сборе, организации, хранении, поиске и распространении информации.
- **Методы**: лингвистические, аналитические, эмпирические, статистические, математическая логика и теория вероятностей, искусственного интеллекта, технологии управления данными.

- **обработка естественного языка -** компьютерное решение задач, связанных с пониманием, анализом, выполнением различных операций над текстами на естественном языке, а также с их генерацией.

Основные понятия

- ***Документ -*** это не юридическая сущность, а содержательно законченная идентифицируемая уникальным образом единица информации, представленная на каком-либо естественном языке.

Представление текстового документа в оцифрованном виде может быть создано с помощью:

- Ввода содержания документа с клавиатуры с использованием какого-либо текстового редактора.
- Сканирования его с бумажного носителя и использования программы распознавания оптических символов (Optical Character Recognition, OCR).
- Генерации текста программным путем распознавателями голоса или какими-либо другими способами.

- Совокупность хранимых в системе документов - ***коллекция документов***.

- Представление информационных потребностей пользователя в форме, воспринимаемой программным обеспечением системы текстового поиска, называется ***пользовательским запросом*** (или для краткости просто запросом).

- Необходимым компонентом содержания пользовательского запроса является описание тех свойств, которыми обладают документы, интересующие пользователя - ***критерий поиска.***
- Хранящиеся в системе документы, которые соответствуют пользовательскому запросу***, называются релевантными.***

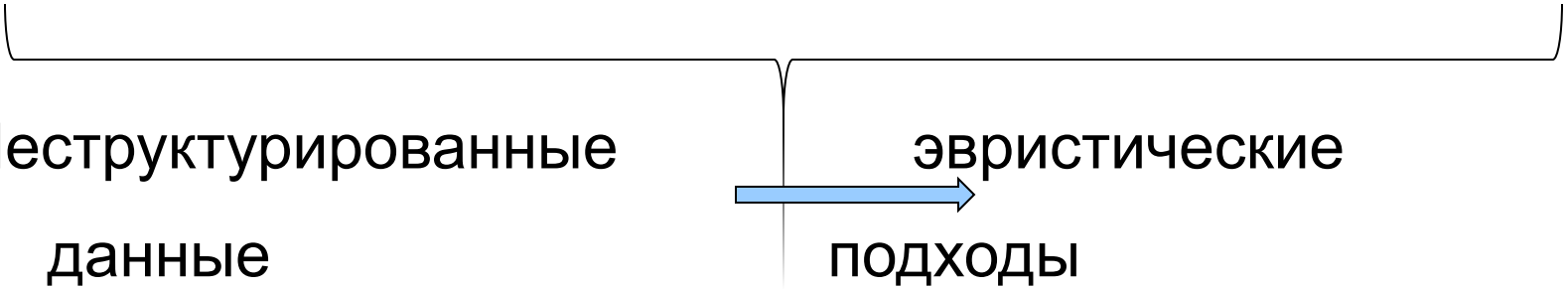
- Некоторые системы текстового поиска выдают пользователю множества документов, полученных в результате обработки запросов, упорядочивая документы по убыванию степени их релевантности – ***ранжирование.***

- в результате обработки пользовательского запроса могут быть найдены документы, не соответствующие информационным потребностям пользователя - ***информационный шум***.

- ***Полнота поиска*** определяет отношение количества релевантных документов, выдаваемых в результате обработки пользовательских запросов, к количеству фактически имеющихся в системе релевантных документов.
- ***Для количественной оценки точности поиска*** может служить доля релевантных документов во множестве результирующих документов запроса.

Принципы текстового поиска
Причины сложности текстового поиска

- Проблемы обработки естественного языка
- Смысловое сопоставление содержания хранимых в системе документов и выраженных на естественном языке пользовательских запросов, оценка степени их близости



Структурированное представление документов

1. Работа со структурированными представлениями документов, формируемыми в результате анализа их текстов, позволяет применять в процессе поиска формализованные методы, основанные на различных эвристических подходах.
2. Производительность системы текстового поиска, анализирующей полные тексты хранимых документов в процессе обработки пользовательских запросов, даже если эта система базируется на очень мощном компьютере, весьма невысокая.

Индексирование документов

- Ассоциированные с документом атрибуты, идентифицирующие документ и/или характеризующие его содержание, называются его индексирующими свойствами.

- На основе индексирующих свойств документов в системе текстового поиска строится ***вспомогательная структура данных (индекс)***, позволяющая по их значениям или по значениям некоторой функции, использующей их в качестве аргументов, эффективным образом (без полного просмотра текстов документов и без полного их перебора) обнаруживать в системных коллекциях документ или документы, которым эти атрибуты соответствуют, и при необходимости осуществлять быстрый доступ к ним.
- Процесс назначения документу указанных атрибутов - ***индексированием документа.***

Способы индексирования

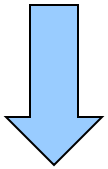
- на основе дескрипторов
- на основе аннотаций, названий или полного текста

Дублинское ядро

это набор элементов метаданных, смысл которых описан вербально и зафиксирован в спецификациях определяющих его стандартов

Стандарты

- DCMI (Dublin Core Metadata Initiative)



- организация национальных стандартов информационных технологий США NISO
- международная организация стандартизации ISO (стандарт ISO: 15836-2003)

Версия DC 1.1. включает 15 элементов метаданных

1. ***Title*** (название ресурса)
2. ***Creator*** (лицо, организация или служба, ответственная за подготовку содержания ресурса)
3. ***Subject*** (тема, обсуждаемая в содержании ресурса)
4. ***Description*** (описание содержания ресурса в свободной форме)
5. ***Publisher*** (лицо, организация или служба, обеспечивающая доступ к ресурсу)
6. ***Contributor*** (другие участники подготовки содержания ресурса, помимо указанного в Creator)
7. ***Date*** (дата создания или предоставления доступа к ресурсу)
8. ***Type*** (жанр, категория или другие характеристики природы ресурса)
9. ***Format*** (характер представления ресурса)
10. ***Identifier*** (точная ссылка на ресурс)
11. ***Source*** (ссылка на источник, из которого продуцирован данный ресурс)
12. ***Language*** (язык представления ресурса)
13. ***Relation*** (ссылка на ресурс, связанный с данным)
14. ***Coverage*** (область пространства, времени и т.д., к которой относится содержание ресурса)
15. ***Rights*** (права интеллектуальной собственности на ресурс и т.п.).

три не вошедшие в указанные официальные стандарты средства:

- Набор из 33 дополнительных и уточняющих элементов метаданных
- Комплект схем кодирования (квалификаторов), каждая из которых определяет множество значений соответствующего элемента DC
- Словарь типов, включающий набор идентификаторов типов возможных значений некоторых элементов DC, указание которых для соответствующих значений позволит адекватно их интерпретировать.

Пользовательские запросы и критерии релевантности

В процессе обработки пользовательского запроса системе необходимо оценивать релевантность очередного рассматриваемого документа

- ***теоретико-множественные критерии***

Функционирование системы текстового поиска

Общие принципы поиска

1. При вводе документа в систему осуществляется индексирование документа и строится его представление, которое будет далее выступать заменителем этого документа в процессе функционирования системы при обработке пользовательских запросов.

2. На основе индексирующих свойств конкретных документов, полученных извне системы или выявленных самой системой путем анализа текстов документов, система формирует и поддерживает индекс для каждой коллекции хранимых в ней документов.

3. *При поступлении в систему пользовательского запроса для него также строится соответствующее представление*.

4. Собственно поиск заключается в том, что каким-либо эффективным образом (не прямым перебором, а обычно с помощью рациональным образом организованного индекса документов коллекции) осуществляется сопоставление представления запроса с представлениями хранимых в системе документов по принятому в системе критерию близости.

Средства лингвистической поддержки

- Словари
- Тезаурусы
- Онтологические спецификации предметной области системы

2. Модели поиска

Модель поиска понимается как сочетание: способа формирования представлений документов; способа формирования представлений поисковых запросов; вида критерия релевантности документов.

Виды

1. Простейшие модели поиска
2. Контекстный поиск
3. Булевская модель
4. Векторные модели

Состояние разработок и новые требования

1. **Структура проблематики текстового поиска:**

- Развитие конкретных моделей поиска
- Методологию проведения экспериментов, тестирования и оценки систем
- Методы и алгоритмы реализации текстового поиска
- Подходы к интеграции технологий текстового поиска и баз данных
- Поиск в среде Веб
- Методы сжатия данных
- Оценку эффективности обработки запросов
- Обработку естественного языка

- Методы классификации и кластеризации текстовых документов
- Приложения информационного поиска в электронных библиотеках
- Глубинный анализ текстов
- Технологии индексирования и поиска мультимедийной информации
- Интерфейсы "человек-компьютер" и т. д.

2. *Развитие функциональных возможностей текстовых систем*

1. *Повышение точности поиска*
2. *Ранжирование результирующих документов запроса*
3. *Обратная связь релевантности*
4. *Автоматическое расширение пользовательских запросов*
5. *Автоматическое индексирование документов*
6. *Мультиязыковой поиск*
7. *Кросс-языковой поиск*
8. *Текстовый поиск в системах баз данных*