Информационный поиск

Лекция 7

Зависимость от структуры

- Реляционные БД vs. Информацинно поисковые системы
- Строгая структура позволяет эффективнее оперировать данными:
  - SQL - select lastname from employees where job_desc like ’invoic%’ (в фамилии)
  - Boolean – invoic* (во всех текстах)

XML

- **XML** (*e**X**tensible **M**arkup **L**anguag e* — расширяемый язык разметки) — рекомендованный Консорциумом Всемирной паутины (W3C) язык разметки. Спецификация XML описывает XML-документы и частично описывает поведение XML-процессоров (программ, читающих XML-документы и обеспечивающих доступ к их содержимому)

Сравнение систем

	Реляционные СУБД	Не структурированный поиск	Структурированный поиск *
Объекты	Кортежи (строки)	Тексты документов	Деревья (листья содержат слова)
Модель	Реляционная модель	Векторная и др.	?
Основная структура данных	Отношение (таблица), индексы (в т.ч. полнотекстовые)	Инвертированный индекс	?
Поддержка запросов	SQL	Произвольные, Булевы	?

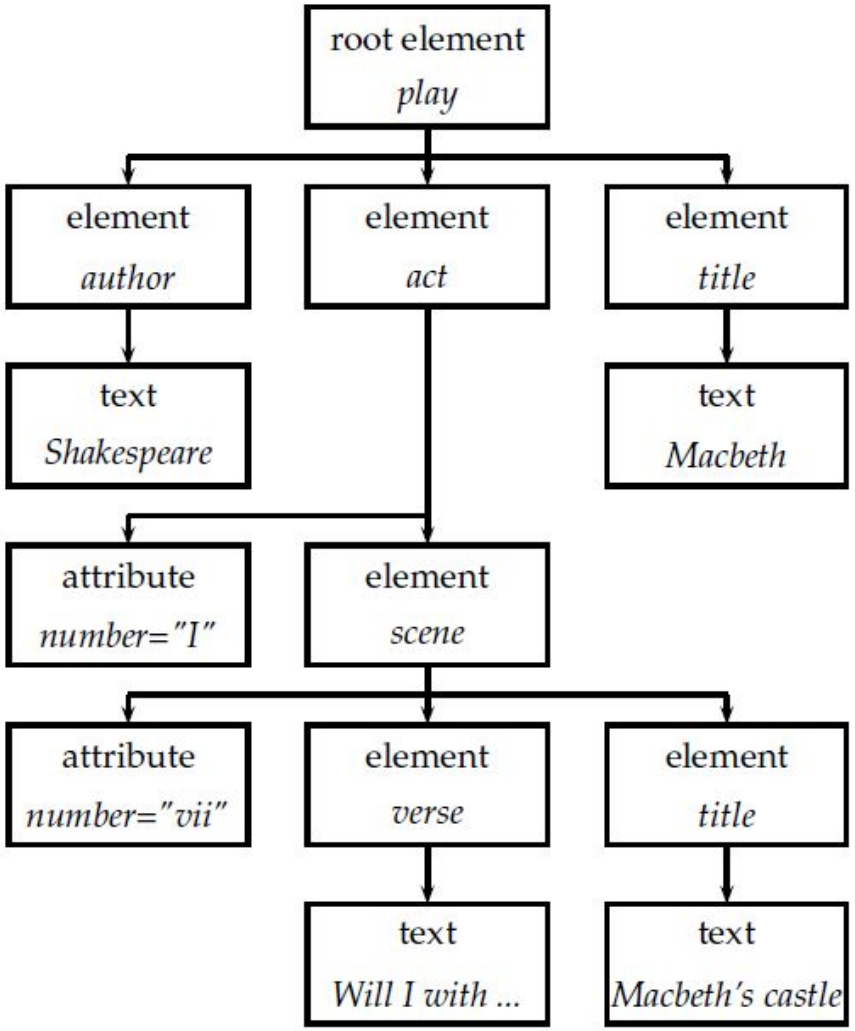
Иногда называют полуструктурированным, чтобы отличить от баз данных

Основные понятия XML

- <**Узел** …> … </**Узел**>
- <Узел **Атрибут1=“Значение” Атрибут2=“Значение”**>…</Узел>
- Узлы могут быть вложенными
- XML DOM (Document Object Model) – представление XML документа в виде дерева узлов с атрибутами

Основные понятия XML

```
<play>
<author>Shakespeare</author>
<title>Macbeth</title>
<act number="I">
<scene number="vii">
<title>Macbeth's castle</title>
<verse>Will I with wine and wassail ...</vers
</scene>
</act>
</play>
```



Основные понятия XML

- Корректность XML документов задается схемой
  - **XML DTD** (Document Data Definition)
  - **XML Schema**
- **XPath** – синтаксис для адресации в XML документах

XPath: Примеры

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<catalog>
  <cd>
    <title>Empire Burlesque</title>
    <artist>Bob Dylan</artist>
    <country>USA</country>
    <company>Columbia</company>
    <price>10.90</price>
    <year>1985</year>
  </cd>
  <cd>
    <title>Hide your heart</title>
    <artist>Bonnie Tyler</artist>
    <country>UK</country>
    <company>CBS Records</company>
    <price>9.90</price>
    <year>1988</year>
  </cd>
</catalog>
```

1. /catalog/cd/price
2. /catalog/cd[0]
3. /catalog/cd/price/text()
4. /catalog/cd[price>10.80]