XML

CISC437/637, Lecture #24

Ben Carterette

Semistructured Data Model

- XML (E**X**tensible **M**arkup **L**anguage) is an alternative to the relational model
  - Rather than tuples of attributes, data is represented using nested tags with content
- Markup:
  - Based on **tags**, e.g. <title>, <author>, <isbn>
- Extensible:
  - Users specify tags and tag semantics; no fixed catalog
- Language:
  - Consists of a set of symbols, a syntax, and a semantics

Advantages of XML

- Semistructured data model
  - Relational data from heterogeneous sources
  - Textual data with tags and links
- Extensible
  - And "self-documenting"
- Flexible exchange format
  - Highly useful for exchanging data between databases/organizations

XML Syntax

- **Tags**: e.g. book, title, author, …
  - Start tag: <book>
  - End tag: </book>
- **Elements**: e.g. <book>…</book>
  - Elements may be nested
  - Empty elements may be present, e.g. <book/>
- **Attributes**: e.g. price
  - <book price="55">…</book>
- **Oids** and **references**: e.g. id="o555"; idref="o555"
  - Defines keys and foreign keys
- An XML document has a single root element
  - It is **well-formed** if all tags have matching, properly nested end tags

XML Semantics

- An XML document is a tree
  - Elements are nodes
  - Leaf nodes are content
- XML allows trees to be *ragged*
  - Relational data viewed as a balanced tree
- XML subtrees need not have identical elements
  - Relational data viewed as isomorphic subtrees connected by a root node

XML Data Typing

- The relational model uses the schema for data typing
- XML uses the **Document Type Definition** (DTD)
  - The DTD describes the valid trees that can occur
  - An XML document that does not conform to the DTD is invalid
    - (Invalid XML can still be well-formed)

