Issues in Accessing XML Data

Fausto Rabitti

ISTI "Alessandro Faedo" - C.N.R.
Via Alfieri, 1 - 56127 Pisa - ITALY

Outline

- What kind of XML data
- XML query language
- Extensions to XML search (Relevance & Ranking)
- XML search in Digital Libraries: the ECD project

What kind of XML data

- XML data on the Web: HTML vs. XML
- XML data from Databases: XML wrappers, XML snapshots
- XML miscellaneous data: Word, Excel, EDI, etc.
- XML Digital Libraries

HTML vs. XML Search

HTML encoding

<UL>
<LI>
R. Goldman, J. McHugh, and J. Widom.
<A href="ftp://db.stanford.edu/pub/papers/xml.ps">
From Semistructured Data to XML: Migrating the Lore Data Model and Query Language
</A>.
Proceedings of the 2nd International Workshop on the Web and Databases (WebDB '99), pages 25-30, Philadelphia, Pennsylvania, June 1999.
<LI>
T. Lahiri, S. Abiteboul, and J. Widom.
<A href="ftp://db.stanford.edu/pub/papers/ozone.ps">
Ozone: Integrating Structured and Semistructured Data
</A>.
Technical Report, Stanford Database Group, October 1998.
</UL>

XML encoding

```
<Publication URL="ftp://db.stanford.edu/pub/papers/xml.ps" Authors="RG JM JW">
<Title>From Semistructured Data to XML: Migrating the Lore Data Model
and Query Language</Title>
<Published>Proceedings of the 2nd International Workshop on the Web
and Databases (WebDB '99)</Published>
<Pages>25-30</Pages>
<Location> <City>Philadelphia</City>
<State>Pennsylvania</State> </Location>
<Date> <Month>June</Month> <Year>1999</Year> </Date>
</Publication>
<Publication URL="ftp://db.stanford.edu/pub/papers/ozone.ps" Authors="TL SA JW">
<Title>Ozone: Integrating Structured and Semistructured Data</Title>
<Published>Technical Report</Published>
<Institution>Stanford University Database Group</Institution>
<Date> <Month>October</Month> <Year>1998</Year> </Date>
</Publication>
<Author ID="SA">S. Abiteboul</Author>
<Author ID="RG">R. Goldman</Author>
<Author ID="TL">T. Lahiri</Author>
<Author ID="JM">J. McHugh</Author>
<Author ID="JW">J. Widom</Author>
```

XML Query Language

- **XQUERY** (W3C draft standard XML query language) distilled from several contenders, including **XML-QL**, **Quilt**, **XQL** and **Lorel**.
- These languages have been developed primarily by members of the database community:
  - These languages provide a declarative way to access elements from XML databases based on various predicates, and rewrite the original XML data into an arbitrary new structure.
  - They generally support (the semistructured equivalent of) the full relational algebra, making them useful for a wide variety of data interchange and transformation applications.

Extended XML Query Languages

- XML query languages (a la DB) do not fully meet the needs of the document management and IR communities.
- Deficit: the lack of support for ranking or weighting of query results based on textual similarity or other relevance metrics.
- It is necessary to extend the XML query languages to encompass both DB and IR functionalities:
  - ELIXIR, extension of XML-QL, supports ranked retrieval based on textual similarity.
  - XIRQL (XQL plus weighting, vague predicates on data types, relevance oriented search, semantic relativism)
  - XXL (flexible XML search language), extension of XML-QL with similarity conditions on elements and their attributes.
  - XXQUERY, extension of XQUERY (with limitations) with similarity conditions on elements and their attributes.

Example – XML objects

<cd_title> The Survivors' Suite
<artist> Keith Jarrett </artist>
<label> ECM </label>
<track> Beginning
<musician> Keith Jarrett
<instruments> piano, recorder, soprano saxophone
</instruments>
</musician>
<musician> Dewey Redman
<instruments> tenor saxophone </instruments>
</musician> ... </track> ...
</cd_title>

<cd title="Full Force" label="ECM">
<artist> Art Ensemble of Chicago </artist>
<track title="Charlie M">
<artist> Lester Bowie
<plays> trumpet </plays>
</artist>
<artist> Roscoe Mitchell
<plays> alto sax, baritone sax, piccolo </plays>
</artist> ... </track> ...
</cd>

Example – XML query

Query: "who plays baritone saxophone on CDs by the Art Ensemble of Chicago"

Q1: ***XQUERY***,
… awfully complex….

Q2: ***XXQUERY with similarity operator ~***

```
For $A in document (http://my.cdcollection.edu/allcds.xml) / ~CD
Where
    $A // artist="Art Ensemble of Chicago“
    and $A // track / ~musician / text() ~ "baritone saxophone"
Return
    $A // track / ~musician
```

XML Search in Digital Libraries (ECD)

- *XML* will play an important *role in Digital Libraries* since such data format can have important impact on the DL components - *metadata, catalogs, and indexes* can make DL information better accessible than general Web information.
- Once multimedia documents are described in XML, it becomes possible to query these sources by means of *semantically more meaningful queries*, taking full advantage of the document hierarchical structure.
- Ontology-based support of *XML Schema management (semantic XML Schema partial integration into trans-schema views)*, to increase the potential *effectiveness/recall* of the query result.

XML Search in Digital Libraries (ECD)

- A query could impose not only lexical, but also *topological constraints* on the documents to be retrieved, so that the mutual nesting of tags is respected: to increase the potential *effectiveness/precision* of the query result.
- A *partial match mechanism* is necessary in processing *partial/imprecise topological constraints*.
- *Query processing efficiency* must be supported by suitable *indexes* (structure based, besides value based) both for *exact access* and for *similarity access*.