

HyREX: Hyper-media Retrieval Engine for XML

Norbert Fuhr

Norbert Gövert

Kai Großjohann

University of Dortmund, Germany
<http://ls6-www.cs.uni-dortmund.de/ir/projects/hyrex/>

1. INTRODUCTION

While XML query languages as proposed by the W3C are targeting towards database-oriented applications and thus do not consider the needs of IR, we developed the query language XIRQL (*XML IR Query Language*) which takes into account the intrinsic imprecision and vagueness of IR. We present HyREX, the *Hyper-media Retrieval Engine for XML*, which implements XIRQL. On the user interface side (HyGate), we present innovative concepts for the presentation of structured retrieval results.

2. XIRQL

The query language XIRQL [1] extends the XPath part of the query language XQuery (proposed standard of the W3C) by the following features:

Weighting and ranking: While XQuery supports Boolean retrieval only, XIRQL allows for weighting document terms as well as query terms. Thus, retrieval results are presented to users as a ranking.

Relevance-oriented search: Traditional IR queries specify only the requested content, but pose no restrictions on the structure of the result. In the case of structured (XML) documents, the IR system should be able to retrieve the most relevant parts of a documents by choosing the most specific element(s) that satisfy the query.

Data types and vague predicates: Since XML allows for a fine-grained markup of elements, there should be the possibility to use special search predicates for different elements of various data types (e.g. person names, dates, names of geographic regions). For each data type, the system must provide appropriate search predicates, most of which should be vague (e.g. phonetic similarity of names, approximate matching of dates, closeness of geographic locations).

Structural relativism: Query languages for XML allow for conditions with respect to the structure of the documents to be retrieved. In order to support uncertainty and vagueness for this type of conditions, appropriate methods ignore the difference between elements and attributes, exploit hierarchies over element

names defined in an ontology, and allow for searching for elements of a specific data type (e.g. search for all elements containing person names).

3. HYREX

The XIRQL language is implemented within the HyREX system. Its system architecture is similar to that of database management systems. Thus, there is a clear separation between the logical and the physical level. At the logical level, XIRQL queries are transformed into a path algebra. A path describes the sequence of document nodes leading from the root of an XML document to a specific element. The path algebra contains operators for manipulating sets of paths that describe intermediate results in query processing.

The connection between the logical and the physical level is formed by the vague predicates, which take a value and/or structure condition as arguments and return a list of paths as result. In order to perform efficient retrieval, appropriate index structures are available at the physical level.

HyREX is Open Source software. It is designed as an extensible IR architecture. The whole system is written in Perl (with time-critical parts in C). For specific applications, new data types can be added to the system, possibly together with new index structures. The current version allows for efficient retrieval of XML collections up to the gigabyte range.

4. HYGATE

HyGate forms the user interface (UI) to HyREX. The development of an UI to an XML IR engine poses a number of new challenges. For query formulation, we present an approach based on the concept of *query by example*; as *example*, either the DTD, the logical structure or the final layout of a specific document can be used. Visualization of retrieval results has to cope with two problems: There may be multiple matches in a single document, and even one match may contain other matches. Here variants of tile bars and tree maps are presented as a visualization of retrieval results.

5. REFERENCES

- [1] N. Fuhr and K. Großjohann. XIRQL: A query language for information retrieval in XML documents. In *Proceedings of the 24th SIGIR Conference*, pages 172–180, New York, 2001. ACM.