

HyREX: Hypermedia Retrieval Engine for XML

Mohammad Abolhassani Norbert Fuhr Norbert Gövert
Kai Großjohann

Chair 6, Department of Computer Science, University of Dortmund

1 Overview

XML¹ is the emerging standard for representing knowledge in almost arbitrary applications. At least almost every kind of knowledge can be represented in XML. For exploring such knowledge, one needs a search engine which is able to let users benefit from all of the concepts with which XML blesses the world.

HyREX is the *Hypermedia Retrieval Engine for XML*. The HyREX project is an ongoing effort (funded as part of other projects like e. g. CARMEN, CYCLADES, and CLASSIX) for developing an information retrieval engine for XML documents. HyREX's main characteristics can be derived from the constituents of its name:

hyper HyREX offers explicit and implicit links to the user. Explicit links are specified within the documents, usually by means of XML linking standards, such as XLink and XPointer. Implicit links are intrinsic to information structures which HyREX derives from XML document collections.

media HyREX offers search facilities for text, but also for other media than text, at least conceptually.

retrieval engine HyREX allows users to explore all kinds of information structures available through XML; besides retrieval in XML documents it allows for browsing and searching the domains of attributes of XML documents as well as schema information given for example by the DTD of a document collection.

XML HyREX allows retrieval under consideration of content *and* structure inherent in XML documents.

In the following we give a brief description of the query language XIRQL which is used within HyREX, and its realization within the system.

2 The XIRQL query language

The current W3C activities for the development of a standard query language for XML (XQuery) are targeting towards database-oriented applications and thus do not consider the needs of IR. In contrast, the Dortmund IR group focuses on document-oriented XML applications, where retrieval must take into account the intrinsic imprecision and vagueness of IR.

For this purpose, the query language XIRQL (XML IR Query Language) is developed, which extends the XPath part of the (proposed standard) query language XQuery by the following features:

Weighting and ranking Whereas XQuery supports Boolean retrieval only, XIRQL allows for weighting document terms as well as query terms. For the former, it is assumed that the weight of a term depends on its context (the definition of these contexts is given as part of an extended DTD). The

¹<http://www.w3c.org/XML/>

underlying probabilistic model treats all term occurrences within the same index node as a single probabilistic event. Query processing produces a Boolean combination of these basic events, for which the correct probabilities can be computed (following the concept of event expressions from probabilistic Datalog).

Relevance-oriented search Traditional IR queries specify only the requested content, but pose no restrictions on the structure of the result. In this case, the IR system should be able to retrieve the most relevant parts of XML 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, technical measurement values, 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 XML query languages allow for conditions w. r. t. 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, searching for elements of a specific data type (e.g. search in all elements containing person names) or by exploiting hierarchies over element names defined in an ontology.

In contrast to XIRQL, XQuery offers additional operators for aggregation and restructuring of results. Further research will focus on appropriate extensions of XIRQL, i. e. probabilistic versions of the corresponding XQuery operators.

3 XIRQL realization within HyREX

The XIRQL language is implemented within the HyREX system. Its 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. After mapping a XIRQL query into a path algebra expression, the query optimization step transforms this expression into an equivalent one which (hopefully) can be processed more efficiently. Since users typically want to see the top ranking elements only, retrieval strategies focusing on these elements will be investigated.

The connection between the logical and the physical level is formed by the vague predicates, which take a value and / or structure condition as argument and return a list of paths as result. In order to perform efficient retrieval, appropriate index structures have to be available at the physical level. Whereas classical inverted lists support value conditions only (indicating occurrence / weights of terms), XIRQL queries also may contain conditions referring to element names and / or indexes as well as to sequence and aggregation of elements. Since inclusion of the necessary information in the inverted list entries will lead to large storage overheads, appropriate compression schemes are investigated.

The development of a user interface to an XML IR engine poses a number of new challenges. HyREX currently supports only a simple Web browser interface where users may enter XIRQL queries and receive ranked lists of answers. For query formulation, several variants based on the concept of query by example are under consideration; as example, either the DTD, the logical structure or the final layout of a specific document can be used. Visualization of results has to cope with the fact that different matches may occur within the same document, where even a match may contain others; here variants of tile bars and tree maps are studied.

HyREX² is Open Source software. The current version allows for efficient retrieval of XML collections up to the gigabyte range.

²<http://ls6-www.informatik.uni-dortmund.de/ir/projects/hyrex/>