Introduction to the Special Issue on INEX

NORBERT FUHR fuhr@uni-duisburg.de
Information Systems, Institute of Informatics and Interactive Systems, University of Duisburg-Essen, Duisburg, 47048 Germany

MOUNIA LALMAS mounia@dcs.qmul.ac.uk
Department of Computer Science, Queen Mary University of London, London, E1 4NS, England, United Kingdom

Abstract. This special issue contains articles describing XML retrieval approaches developed and evaluated during the second year of INEX, the evaluation initiative for XML retrieval.

1. XML retrieval and evaluation

Today’s content is increasingly a mixture of text, multimedia, and metadata. One way to format this mixed content is according to the adopted W3C standard for information repositories and exchanges, the so-called eXtensible Markup Language (XML).

The increasing use of XML in scientific data repositories, Digital Libraries and on the Web, has brought about an explosion in the development of XML systems, and in particular systems to store and access XML content. Whereas many of today’s systems still treat documents as single large (text) blocks, XML offers the opportunity to exploit the internal structure of documents in order to allow for more precise access by giving more specific answers. Providing effective access to XML-based content is therefore a key issue.

Effective access to XML-based content is what XML retrieval research is about. XML retrieval systems aim to exploit the logical structure of documents, which is explicitly represented by the XML markup, to retrieve document components (the so-called XML elements) instead of whole documents in response to a user’s query. Implementing this more focused retrieval paradigm means that an XML retrieval system needs not only to find relevant information in the XML documents, but also determine the appropriate level of component granularity to return to the user. Evaluating how good these systems are, hence, requires test-beds where the evaluation paradigms are provided according to criteria that take into account the imposed structural aspects.

INEX, the Evaluation Initiative for XML Retrieval, is an initiative currently supported by DELOS. INEX was set up at the beginning of 2002 with the aim to establish infrastructures, XML test collections and appropriate measurements for evaluating XML retrieval approaches. INEX starts every year in April and concludes with a workshop, held a Schloss Dagstuhl (Germany), in December.
2. The INEX test-bed

The INEX document collection is made up of the full-texts, marked up in XML, of 12,107 articles of the IEEE Computer Society’s publications from 12 magazines and 6 transactions, covering the period of 1995–2002, and totalling 494 megabytes in size. The collection contains scientific articles of varying length. On average an article contains 1,532 XML elements, where the average depth of an element is 6.9. Overall, the collection contains over eight millions XML elements of varying granularity (from table entries to paragraphs, sub-sections, sections and articles, each representing a potential answer to a query).

To consider the additional functionality introduced by the use of XML query languages, which allows the specification of structural query conditions, INEX defined two types of topics:

- Content-only (CO) queries are standard information retrieval retrieval similar to those used in TREC. Given such a query, the goal of an XML retrieval system is to retrieve the most specific XML element(s) answering the query in a satisfying way. Thus, a system should e.g. not return a complete article where a section or even a paragraph of the same document may also be sufficient.
- Content and structure (CAS) queries contain conditions referring both to content and structure of the requested answer elements. A query condition may refer to the content of specific elements (e.g. the elements to be returned must contain a section about a particular topic). Furthermore, the query may specify the type of the requested answer elements (e.g. sections should be retrieved). The query language defined for this purpose is a variant of XPath 1.0.

As in TREC, an INEX topic consists of the standard title, description and narrative fields. Keywords used in the topic creation phase are also included. The INEX topics were created by the participating institutions using their own XML retrieval systems or the system provided by the INEX organisers.

Two retrieval tasks were performed by the participating groups, both ad-hoc retrieval of XML documents, for CO and CAS topics, respectively. Retrieval runs for each task were then submitted to the INEX organisers.

Like the topics, the assessments have been derived in a collaborative effort. For each topic, the retrieval runs from the participants’ submissions have been collected into pools using the pooling method. The assessments pools were assigned then to participants; either to the original authors of the topic when this was possible, or on a voluntary basis, to groups with expertise in the topic’s subject area. Each group was responsible for about two topics.

For the construction of the relevance assessments, INEX employed two relevance dimensions: exhaustivity and specificity. Exhaustivity is defined as a measure of how exhaustively an XML element discusses the topic of request, while specificity is defined as a measure of how focused the element is on the topic of request (i.e. discusses no other, irrelevant topics). Both dimensions are measured on four-point scales with degrees of highly (3), fairly (2), marginally (1), and not (0) exhaustive/specific. The motivation for the use of