Second Edition of the "XML and Information Retrieval" Workshop Held at SIGIR'2002, Tampere, Finland, Aug 15th, 2002

Ricardo Baeza-Yates, University of Chile, Chile
Norbert Fuhr, University of Dortmund, Germany
Yoelle S. Maarek, IBM Research Lab in Haifa, Israel

Introduction

The previous workshop on "XML and Information Retrieval" was held in the context of SIGIR'2000 (Athens, Greece) and showed that there is a serious interest in managing semi-structured data from an IR (i.e., unstructured) perspective rather than from the dominating database (i.e., structured) perspective. As a direct outcome of the workshop, a special JASIS issue on XML and IR was edited and published in March 2002. The topic is still in its prime both in the Academia, as evidenced by XML related papers presented at SIGIR'2001, and in the business world with more search engines adding "XML support" as a requirement.

The purpose of this workshop was to continue the effort of applying an "IR approach" to XML search and retrieval. We believe that this approach is crucial to extending the role of XML from pure data exchange in business-to-business applications to actual information exchange in end-users facing and Knowledge Management applications. As a side issue, we also intended to ask the IR community how it should relate, if at all, to the XQuery initiative at W3C. The workshop included 3 presentation sessions, for a total of 8 reviewed papers, an invited talk and a panel. The workshop attracted an audience of more than 30 people, who were encouraged to interact with the speakers as well as the audience.

XML Query and Retrieval

The first session of the workshop included 3 papers. Torsten Grabs, from ETH in Zurich, opened the morning with his work co-authored with Hans-Jörg Schek on *"Generating Vector Spaces on the Fly for Flexible XML Retrieval"*. Torsten first discussed the shortcomings of conventional IR for XML collections, which are mostly due to the fact that IR assumes flat document structures and that indexing granularity is restricted to complete documents or predefined fields, thus term statistics on which ranking is based refer to a very precise scope. Since XML has a heterogeneous form, queries might have very different scopes and term statistics should reflect the scope of the query. This scope will typically be a DOM subtree, or in the case of a multi-category query, the statistics for the term weights should be derived on the fly from the new universe, or vector space, of the query. Törsten presented the flexible IR model for single-category, multi-category and nested retrieval where the basic index and statistics data are integrated on the fly at retrieval time depending on the scope of the query. Preliminary results for different categories of flat documents showed that the dynamic integration required less than 30% overhead with up to 16 different categories. The authors have good hope that they will obtain also reasonable results on arbitrary XML collections.

The second paper entitled "*An Extension of the Vector Space Model for Querying XML Documents via XML Fragments*" also discussed a variation of the vector space, which concentrated on a new way of expressing queries rather than on their scope. This work was presented by Yoelle Maarek, and was co-authored with, David Carmel and Yosi Mass, from the IBM Research Lab in Haifa, and Nadav Efraty and Gad Landau from Haifa University. Yoelle argued that, when the basic user's need in querying an XML collection is an information or discovery need rather than a management need, the simplest approach is to follow an IR model and express this "fuzzy" need as an XML fragment that can easily compare to XML documents. This approach of expressing queries of the same form of the objects being searched thus imitates the full-text search approach where queries are expressed as free-text rather than formal Boolean queries. Yoelle then presented a novel measure of similarity between XML fragments and XML documents that is an augmentation of the classical cosine measure of similarity, where not only the textual part of the XML structure is compared but also the context of its occurrence (which can be expressed as a path in the DOM tree). She also introduced a measure of similarity between such paths, which allowed users to be approximate in specifying not only the free-text part but also the various contexts. She finally showed some examples of querying an XML collection of car reviews via XML fragments, and ranked results associated with these examples.

The last presentation of this session was given by Kai Großjohann from the University of Dortmund, and co-authored with Norbert Fuhr, Daniel Effing and Sascha Kriewel. Their work entitled "*Query Formulation and Result Visualization for XML Retrieval*" is a follow-up of the XIRQL work presented at the previous edition of this workshop at SIGIR'2000. Kai reminded the key principles of XIRQ. Queries consist of a set of atomic query conditions, which are combined either via Boolean or structural combination. Kai argued that the primitives needed for dealing with XML need to be reflected both in the query formulation and the results visualization. He introduced a novel interface for expressing queries in an XPath-like form. The query by example interface lets users pick a term (or value) in an XML document, and obtain all the preferred path expressions for that term location, with the possibility to filter, or expose the commonality between paths so as to allow for structural combinations. The second contribution of this work is a method for visualizing results that shows in a unified form fragments possibly originating from different documents and yet keep the tree structure of the XML documents. Kai used a variation of the well-known treemap method, called "partial treemap" that omits nodes that are not a retrieved node or an ancestor of a retrieved node. He concluded by discussing experiments that compared text, tilebars and partial treemaps.

XML Stores

The second session dealt with XML stores and was opened by a paper on *"Knowledge Discovery from Mixed-model XML documents"* by Junji Tomita, Tetsuo Ikeda, Tamio Kihara and Tetsuji Satoh from NTT Cyberspace Laboratories. Tomita-san, who presented this work, proposed a comprehensive analysis method for analyzing both text and meta-data encoded in mixed-model XML documents. He proposed to store the textual part in an IR system (or text database as referred to by the authors) and the meta-data in a relational database so as to provide rich text and metadata handling operation. The suggested analysis method involves 3 steps: specifying entity that should be analyzed (via XPath), specifying the "viewpoint" (how it should be analyzed), and classifying the viewpoint into one of two classes text or metadata, according to the operations to be conducted on it (which include searching, clustering, summarizing, visualizing, etc.). A preliminary evaluation was conducted on a collection of Japanese patents, and a live demo of the system was given to the workshop audience.

The second[1] paper of this session was *"Index Infrastructure for an XML repository"* by Alain Azagury, Michael Factor, Naama Kraus, Irit Loy and Benjamin (Benny) Mandler from the IBM Research Lab in Haifa. Benny presented this work, which was a follow-up of a previous work on the XMLFS introduced at the previous edition of this workshop at SIGIR'2000. Benny described a generic infrastructure for XML repositories that coded both textual and structural information into a single repository rather than in 2 distinct ones like in the previous paper of this session. Support for XML is achieved via a simple extension of the classical trie-based lexicon, and associated inverted index, used by many IR systems. The extension consists of decorating each word appearing in the full-text part of XML documents (e.g., "Paul") with its context, where the context is a unique id (e.g., "4") that represents the XPath (e.g., "/name/first") of a word in question. Thus in our example instead of storing simply "Paul" in the lexicon, the new entity "Paul#4" is stored and points to the postings of the documents that contain Paul under "name/first" rather than anywhere in the document. Benny then explained how this store is generic enough to support a variety of operations from regular free-text, to context sensitive, relational, partial and composite operations to cite just a few.

Modeling Semi-Structured Data

The third session of the workshop included 2 papers. The first one entitled *"Processing Queries with Metrical Constraints in XML Based IR Systems"* by Shmuel Tomi Klein from Bar-Ilan University. Tomi explained how he revisited, in the context of XML, a previous work of his on the computation of word proximity across annotations on full-text. XML documents are more complex than annotated text for 2 reasons: annotations have by definition an inferior status than text, which cannot be assumed in XML, and annotations represent only a second layer of metadata while the metadata encoded in XML is a full tree. Tomi then offered a formal definition for expression metrical

[1] Due to laptop problems, this presentation was actually given right after the lunch break in place of Tomi Klein's presentation, and Tomi kindly accepted to replace Benny on the fly. Yet for the sake of the structure (no XML pun intended:-), we keep ,in this report , the original order of presentations.

constraints on queries. He explained the need to define the coordinates of tokens in the DOM tree, to label the DOM tree nodes and within nodes, define a flat representation. Word proximity constraints are then allowed exclusively across elements that are either siblings (left to right and right to left), or are descendants (in top down order only). Tomi then discussed implementation issues and concluded by insisting that the presented work was theoretical in nature and needed to be validated by experimentations in order to be of practical value.

The final paper of this session presented *"A Bayesian Network Model for Document Retrieval in a Hierarchically Structured Collec*tion". It was co-authored by Benjamin Piwowarski and Patrick Gallinari, from University Paris 6. Benjamin started his presentation by stating the goal of this work, which is to support content only and content and structure queries. He proposed a model to compute and combine the scores of document parts, which is based on Bayesian networks and makes use of a learning machine method for learning the model parameters when faced to incomplete data. More precisely, the Bayesian Network learns its conditional probability tables via the Estimation Maximization method. While this work was first targeted at the specific task of retrieving Web pages from a hierarchically structured Web site, it could easily be extended to the retrieval of documents from an XML corpus by mapping the document structure into a Bayesian network, the same way a Web collection can be mapped into a structure. Benjamin concluded by describing the experiments they conducted as well as mentioning the few current limitations of the model, which include the fact that only a special case of structure was considered and the fact that textual information was not included in the network.

Invited talk: INEX

The invited talk of the workshop dealt with the INEX evaluation initiative, to which many of the workshop attendees participated. An abstract of this talk entitled *"INEX: Initiative for the Evaluation of XML Retrieval"* by Norbert Fuhr, Norbert Gövert from the University of Dortmund, and Gabriella Kazai and Mounia Lalmas, from the Queen Mary University of London is included in the Workshop Notes.

Gabriella started by stating the aim of INEX, which is to support research within the XML retrieval community by providing a large XML test collection, evaluation measures and, most of all a forum to compare results. One key point of the initiative is that it has a document centric view of XML for relevance-oriented retrieval. The test collection consists of 3 parts: a set of XML documents, a set of queries and the associated relevance assessments. The XML documents were donated by the IEEE Computer Society and consist of more than 12000 articles representing about 500MB. The query set consists of 60 queries, half of them were "content-and-structure" (CAS) queries, and half of them were "content-only" (CO) queries. Queries were selected from a larger set provided by the INEX participants. At the time of the workshop, relevance assessments, which were also to be provided by the INEX participants, had not been conducted yet. The audience strongly supported the initiative, and all agreed that it is a "must" for promoting research in the field. Two concerns were expressed though with respect to the form of the

evaluation: Djoerd Hiemstra regretted that guidelines on how assessments would be conducted had not been published ahead of time, and David Carmel requested that participants be at least educated on the guidelines before the assessment starts. The INEX workshop will be held in Dagstuhl, Germany in Dec (9-11), 2002.

Panel

The workshop ended with a panel to which participated, Ross Wilkinson, from CSIRO, Ian Soboroff, from NIST, David Hawking from CSIRO and Djoerd Hiemstra from the University of Twente. Yoelle Maarek, from IBM Research Lab in Haifa acted as moderator and initiated the debate by asking the panel whether XML retrieval (and this very workshop) make sense at all, and if they do, how would the conjecture that it would provide better results than full-text search would be proven. Ross Wilkinson proposed the following answer: "Does XML retrieval make sense? The answer is not 'yes or no', but rather absolutely some time!" which provoked a huge laugh in the audience. More seriously, Ross believes that XML search will work when it captures this reduction and when there is an agreement on what it means. He does not believe on a general technique, and is pretty skeptical on the semantic Web simply because we will never agree on a common structure. He added that the only way to prove the conjecture is to go back to the people to verify whether in the end, they can perform the task at hand more effectively. Ian Soboroff agreed with Ross, as he believes that XML retrieval will work in the proper settings. David Hawking offered that IR should stand for "Inexact Retrieval", and he claimed that there is nothing really special about XML, "it's only annotation". Yet in the case of multi-source collections, retrieval starts to be really complicated unless we all agree on DTDs. David then mentioned Andrei Broder's classification of Web users' needs into informational, navigational and transactional. The users' needs, in the context of XML, have also to be identified, in order to answer the first question of what to retrieve. He concluded by a rhetorical question of whether XML document retrieval systems should pay a certain price to support DB operations? David answered his own question with a categorical "No!" and concluded by provocatively telling the audience "XQuery who cares?" which earned him enthusiastic applauses from the audience. In trying to address the conjecture question, Djoerd stated that he did not know what the killer application would be (which would prove conjecture), yet XML retrieval is crucial and those who argue are probably the same who said, 20 years ago, "who cares about ranked retrieval!" He agreed that it will be very hard to prove it but that should not discourage us. The debate then became more interactive with the audience. Examples of niches or domains (such as the full Finish health records) and suggestions on how to find the killer application were offered.

Conclusion

The workshop ended with the panel on a very positive tone, where most felt encouraged to continue their research. The informal atmosphere triggered many debates, and part of the audience even stayed after the conclusion of the workshop for more discussions. The workshop was considered a success, and indeed obtained the best evaluation by SIGIR'2002 participants.