XML and Information Retrieval: a SIGIR 2000 Workshop

David Carmel, Yoelle Maarek, Aya Soffer

IBM Research Lab in Haifa

Introduction

XML - the eXtensible Markup Language has recently emerged as a new standard for data representation and exchange on the Internet. It is believed that it will become a universal format for data exchange on the Web and that in the near future we will find vast amounts of documents in XML format on the Web. As a result, it has become crucial to address the question of how large collections of XML documents can be sorted and retrieved efficiently and effectively.

To date, most work on storing, indexing, querying, and searching documents in XML has stemmed from the database community's work on semi-structured data. An alternative approach, that has received less attention to date, is to view XML documents as a collection of text documents with additional tags and relations between these tags. IR techniques have traditionally been applied to search large sets of textual data and should thus be extended to encode the structure and semantics inherent in XML documents. Integrating IR and XML search techniques will enable more sophisticated search on the structure as well as the content of these documents, while leveraging the success of IR techniques in document similarity ranking and keyword search.

The SIGIR workshop on XML and information retrieval was held July 28th, in Athens Greece. The goal of the workshop was to bring together researchers and practitioners interested in XML and IR to discuss and define the most relevant topics in the relation between these two technologies, present recent results, and propose future directions for research. The topics for discussion included:

- How to extend IR technologies to search XML documents
- How to integrate XML structure in IR indexing structures
- How to query XML documents both on content and structure
- How to introduce the semantics inherent in XML into the search process
- How to adopt database indexing techniques in an IR framework

The opening session of the workshop consisted of a survey of search engines for XML documents. This was followed by three technical sessions: query languages, retrieval algorithms, and IR systems for

XML documents. The final talk of the day, "Searching Annotated Language Resources in XML", by Nancy Ide was given from the perspective of potential users of XML search systems and opened many topics for discussion. The workshop was concluded with a panel discussion where the panelists outlined their vision of the future of XML search.

XML Search Engine Survey

The first talk of the workshop was a survey of search engines for XML documents given by Robert Luk from the Hong Kong Polytechnic University. Luk classified the main existing search techniques for XML documents to database oriented approaches, IR oriented approaches, and hybrid approaches. In the database oriented approach, data derived from databases is converted to XML documents which are rendered using XSL stylesheets. The advantage of this approach is that data relations, constraints and integrity can be modeled and checked, standard database engines can be used, and the query language is similar to the well known standard database query language (SQL). The disadvantage is that importing XML data into a database is difficult and furthermore it is very hard to handle changes in schemas. In the IR approach, basic IR techniques are directly applied to the processing and retrieval of XML documents, each of which is considered a text document with additional markup overheads. Several methods have been suggested to deal with the XML tags. These include simply discarding all tags; selecting some important tags and inserting them into the index; indexing all of the tags as if they were the index terms. In the hybrid approach, a more accurate method is used to specify the nesting elements using Xpaths. The basic idea is that inverted lists of all terms and their associated Xpaths are indexed. This is likely to give more precise search results since the Xpaths specification limits the possible matches. The drawbacks are that some storage overhead is paid for this flexibility, and that the user must have some additional knowledge about Xpaths for more precise results.

Query languages

The first technical session consisted of two papers on the topic of query languages for XML search. Kai Grossjohann of Dortmund University, Germany, presented XIRQL, an Extension of XQL for Information Retrieval. Grossjohann first described the XQL query language. He claimed that while XQL supports retrieval with respect to the tree structure of the XML documents, it lacks some important IR search criteria such as:

- no weighting of search results,
- no vague predicates to measure similarity with respect to sounding, classification etc.,
- no relevance oriented search, where only relevant nodes in the XML tree should be returned,
- no semantic relativism among different XML tags,

Grossjohann then presented XIRQL which is an extension of XQL that supports these missing features. XIRQL is based on pDatalog, a variant of Datalog where facts as well as rules can have attached probabilities. pDtalog supports intentional semantics and independent events, and each document is transformed into a set of EDB predicates. A XIRQL query is transformed

into IDB predicates and a pDatalog query. The XIRQL specification adds operators that enable the use of vague predicates, performing relevance-ortiented search, and abstraction of data types and tags.

One question raised by the participants was that XIRQL seems like a very complicated language for the average user. The response was that it is not meant to be used by and end user. It should be used either by an application or via some dedicated GUI. Another question raised was what is the result of an XIRQL query? Is it the entire document, a part of the XML tree, or perhaps a graph. There was no good answer for this and the consensus was that this is an open question.

The second paper of this session was presented by Albano of Dipartimento di Informatica University di Pisa who presented a type system for querying XML documents. Albano described a type approach to process and to query XML documents. A type represents a collection of XML documents sharing the same structural properties. This approach exploits the type information for checking query correctness, for providing the user with information about the structure of the data, and for performing some query optimization. Algorithms for checking conformity and query correctness based on the document type were described.

Albano was asked how this model can be compared to XML-Schema. He said that it is different it is a theoretical system. An additional concern was that this model does not address traditional IR search criteria such as weighting, vagueness, etc.

Retrieval algorithms

The second technical session focused on retrieval algorithms for XML search. Three papers were presented. First, Yoshihiko Hayashi from NTT Cyberspace Laboratories, Japan presented a technique that his group has developed for searching text-rich XML documents with relevance ranking. This system uses a format file which describes the search fields (some of the XML tags defined to be searchable) and the index type for each tag. The indexer creates a separate index for each search field using traditional IR techniques. The query engine submits the user query to the relevant indexes according to the format file. Results are combined and ranked. The index size is about 20% of the original data and indexing time is 0.01 sec/document. While this system seems to handle predefined simple DTD's well, there was some concern among participants that having a separate index for each search field makes this applicable only to documents with a small number of fields. Furthermore, the fact that the tree structure is part of the index, makes the method inflexible in terms of handling updates in the XML schema.

Ricardo Baeza-Yates from the Universidad de Chile described how the proximal nodes data structure can be used to support XQL queries. Baeza-Yates showed how XQL can be implemented using the Proximal nodes model. This model specifies a fully compositional language with three type of operators: (1) text pattern matching; (2) structural retrieval; (3) combining results. The main idea behind efficiently computing these operations is to first search on the content and then evaluate the structural part. Two separate indices are used for text and for structure. These indices support operations of type (1) and (2). To ensure that operations of type (3) can be processed efficiently, only operations that relate "nearby" nodes are allowed. Nearby nodes are those whose segments are more or less proximal. This model can be efficiently implemented, being linear for most

operations and in all practical cases. The proximal nodes model permits any operation in which the fact that a node belongs or not to the final result can be determined by the identity and text position of itself and of nodes in the operand which are proximal to it. The question of what constitutes a useful query result was raised once more in the context of this paper. Another issue that was raised was how can you rank documents based on the structure. What is a good metric to evaluate the distance between two different structures?

Finally, Torsten Schlieder of Freie Universität Berlin presented a method for querying XML data using approximate tree embedding. Schlieder presented an approach for querying XML documents by reducing the problem to the unordered tree inclusion problem. In this approach the goal is to approximately embed the query tree into the data tree such that the labels and the ancestorship of the nodes are preserved. There is no polynomial time algorithm for the approximate tree embedding problem. However, the authors, described a new algorithm that is still exponential in the worst case but in typical cases it takes sublinear time with respect to the database size. Schlieder uses mainly similarity in terms of structure for ranking. One issue raised was how can this system also use text similarity to rank results. Another issue brought up by this paper (and also throughout the workshop) was what are good sources of XML data for experimentation.

IR systems for XML documents

The last technical session featured three presentation of existing commercial and prototype systems for archiving and searching XML documents. Two of these presentations also included demonstrations of the systems. The first system was presented by Robert Lord from XYZFind Corporation. XYZFind is a system for structured information retrieval using XML. Lord described XYZFind, a search system for XML documents and discussed its main ideas:

- Techniques for exploiting semantically structured XML to increase precision and recall.
- A user interface model which allows end users to engage the system in a dialogue to help create structured queries.
- Extension to the classic inverted index to support structured IR.

Lord demonstrated some of the features of the system. In particular, he demonstrated a dynamic user interface in which the valid values of fields can be determined based on the the XML structure as well as the values in the database. Many questions were raised concerning the assumptions made by the system (such as what is a valid price for a car) and whether this is a reasonable model. Lord's response was that in general their experience is that these assumptions and use of the document structure can enhance the user's search experience and that any help a user can get in posing queries is worthwhile.

The second system, Xyleme, was presented by Vincent Aguilera from INRIA, France. The Xyleme project proposes to build and study a dynamic World wide XML warehouse which is capable of storing and searching XML data. In this work, the authors focused on query processing aspects of Xyleme. The language they use is an extension of OQL and provides a mix of database and IR characteristics.

Finally, Benjamin Mandler from the IBM

Research Lab in Haifa presented XMLFS: An XML-Aware File System. XMLFS combines the interface of file systems with the organizational power of IR to provide a repository for XML documents. As such, it provides a solution for organizing, searching and browsing collections of XML documents. The key difference between XMLFS and traditional hierarchical file systems, is that instead of showing files organized in a static directory structure, XMLFS shows files organized in in a dynamic hierarchy which is constructed on the fly. The dynamic structure is based upon the file's content, were the content is extracted from the attributes defined by the XML structure. A directory path in XMLFS is a sequence of attributes and values, and the contents of a directory are all XML documents which have the attributes/values named in the path.

Searching Annotated Language Resources in XML - An Application

The final talk of the day was presented by Nancy Ide from Vassar College, USA. Nancy's talk gave the perspective of potential users of XML search systems who would like to search annotated language resources in XML. Ide described methods for annotating language resources. In particular, she described XCES (the Corpus Encoding Standard in XML) that is being developed by the Department of Computer Science, Vassar College, and Equipe Langue et Dialogue, LORIA/CNRS. XCES provides the mechanism to mark up structural and linguistic information in large text corpora, discourse/dialogue, lexicons, and speech. It is based on the CES standard which specifies a minimal encoding level that corpora must achieve to be considered standardized in terms of descriptive representation (marking of structural and typographic information) as well as general architecture (so as to be maximally suited for use in a text database).

A question was raised whether it is currently possible to take advantage of the rich annotations for search purposes. Ide indicated that there is no such facility and that from a linguistics perspective it would be valuable to search both on content and structure.

Panel Session

At the close of the workshop, a panel discussion was held where the panelists outlined their vision of the future of XML search. The panelists were David Hawking of CSIRO Australia, Robert Lord of XYZFind corporation, and Ricardo Baeza-Yates of the Universidad de Chile.

One of the main issues that surfaced in the panel was whether XML will in fact be the lingua franca of the Web, or will it be restricted to B2B applications. It is crucial to know who will be the end users of XML data, novice users who need user-friendly search paradigms, applications that need to access databases, or both. The usage will determine what features are necessary for the XML search paradigm.

Conclusion

We are still lacking large XML repositories and furthermore it is not yet clear which applications will adopt XML as a tool for rich text representation. We cannot hope for miracles, XML data is as informative as the content the authors put in it. For instance, WML which is now widely used for Web content geared to the wireless community, uses the markup features of XML mostly for layout oriented information. This limited use of XML does not provide information about

the content of the document that could be used to facilitate enhanced search. It may thus be too early to resolve issues such as the desired query interface and functionality of XML search systems. It is, however, crucial to work on the infrastructure of such search facilities and demonstrate the potential benefit of rich annotations. Once users are aware of these benefits, they may be willing to invest in creating rich content using XML. We are basically facing the classic chicken and egg problem here. Nobody is going to invest in creating rich XML content if there is no way to take advantage of the annotations, and nobody wants to invest in creating systems that take advantage of the annotations if there is no such data.

From a research perspective, there are many important issues to address. First, it is important to work on generating sources of XML data for evaluation purposes as well as develop evaluation methods. Second, we must address questions such as whether there can be one generic XML search paradigm whatever DTD/schema is used, what constitutes a valid response to an XML query, how can we rank results based on both content and structure, what are efficient data structures for supporting such search, etc.

In conclusion, the topic of XML and information is clearly in its prime. The workshop generated a great amount of interest, and attracted more than 50 participants. We have received numerous requests for the working notes of the workshop. As a result of the success of the workshop, the co-chairs along with Ricardo Baeza-Yates will be guest editors of a special topic issue of the Journal of the American Society for Information Science (JASIS) on the topic of XML and Information Retrieval that is scheduled to be published in mid 2001.

Acknowledgments: We would like to thank Sue Dumais for her quest to bridge the gap between the IR and Web communities that made this workshop possible, as well as all of the workshop participants for making this workshop a success and keeping their good spirit even during the power outage.

For more information about the XML and Information Retrieval workshop, visit the workshop Web page at http://www.haifa.il.ibm.com/sigir00-xml/ or contact any of the organizers: David Carmel **carmel@il.ibm.com**, Yoelle Maarek **yoelle@il.ibm.com**, Aya Soffer **ayas@il.ibm.com**