Third Edition of the "XML and Information Retrieval" Workshop
First Workshop on Integration of IR and DB (WIRD)
Jointly held at SIGIR'2004, Sheffield, UK, July 29th, 2004

Ricardo Baeza-Yates, University of Chile, Chile
Yoelle S. Maarek, IBM Haifa Research Lab, Israel
Thomas Roelleke, Queen Mary University of London, UK
Arjen P. de Vries, CWI, Amsterdam, The Netherlands

Introduction

The morning session was dedicated to the third edition in the series of XML and Information Retrieval workshops that were held at SIGIR'2000 (Athens, Greece, see SIGIR Forum Fall 2000 issue) and SIGIR'2002 (Tampere, Finland, see SIGIR Forum Fall 2002 issue). The goal of the workshop, co-chaired by Baeza-Yates and Maarek, was to complement the INEX (Initiative for the Evaluation of XML Retrieval) meetings that have been organized for the last two years, by providing researchers a useful forum for discussing (before implementing) and evaluating their models at INEX in the second half of the year. Our intent was twofold: first encourage the exchanges of ideas between researchers who are now active in this "sub-field" and, second, attract new interests. Our focus, like in previous editions of the workshop, was to address issues related to the application of IR methods to XML data for querying, retrieval, navigating, etc. We have gone a long way since the first edition in 2000, when XML was entirely dominated by the DB community. However, it seems that the expected breakthrough has not occurred yet, and it is not clear whether it is for lack of XML data, of appropriate technology, or simply of real needs in the marketplace.

The afternoon session was dedicated to the first Workshop on the Integration of Information Retrieval and Databases (WIRD'04). The purpose of this workshop has been to bring together researchers of the database (DB) and information retrieval (IR) fields, facilitating exchange on the progress in developing and applying integrated IR+DB approaches (or, naturally, DB+IR solutions). The three papers selected for the workshop covered different aspects of data abstraction and DB usage for IR, illustrating how a different perspective on the integration leads to quite different views: a high level of abstraction improves the development of advanced retrieval systems, following by a (layered) database architecture that may enable efficient processing of complex XML-IR queries, or, where structured data could be used for query expansion in a more traditional information retrieval setting. Also, it seems safe to expect to see several of the attendees jumping on the general problem of 'enterprise search'. We believe that the presentations approaching the problem from different angles, and the ample opportunity for discussion, as well as the attendees' inspiration triggered by the morning excellent keynote really have made the workshop a forum for exchange on future DB+IR technology!

Invited talk: David Hawking, "*A Panoptic View on Databases, XML and IR*"

David Hawking from CSIRO in Canberra, Australia, was the invited speaker of both workshops. He discussed his view on Databases, XML and IR from the perspective of an enterprise search engine, Panoptic, that started around 1999 and has now turned commercial, being deployed at around 35 sites. While this specific search engine is clearly not a major player in the enterprise world, it presents interesting features. The latter were specifically designed for enterprise customers from various horizons (i.e., oil exploration companies), who need to handle a mix of more or less structured documents and meta-data, originating from heterogeneous sources, ranging from email and regular documents to spreadsheets and DB records. Some of the issues that Hawking then rose included the handling of heterogeneity, the distinction between the notions of value and of relevance in search results, indexing DB entries as opposed to the Web pages derived from them, and more generally the boundaries between IR and DB.

Hawking then illustrated his purpose with technical examples on how Panoptic supports semi-structural queries, relying on meta-data classes. The basic idea is to map meta-data classes to single letters of the alphabet, hence a query for "Baeza-Yates" in an author field (be it specified in XML or originating from a DB table) will be expressed as

`"a:Baeza-Yates".`

Hence, the number of classes is limited to 26 letters, which according to Hawking is in most cases largely sufficient. He made the point that this approach raises the lowest common denominator for search over combination of web pages, email, xml, db (over text), through a single mapping instruction, which is feasible in an enterprise environment. He gave some more details on the query language, which is not Boolean (often found too complex by many). An implicit "and" is assumed between all query terms and alternatives can be specified via square brackets. In addition, wildcard and special operators are supported. An example of valid query is given below:

```
"workshop on db and ir" [a:Maarek a:Marek] p:ibm d>feb
```

where the square brackets indicate alternative spellings and `d` is a reserved token for dates. In terms of implementation, Panoptic uses db2xml plus a JDBC bridge to convert DB records to XML files, and records/fields to be indexed are selected via SQL queries.

Hawking concluded his talk by raising the question of how far an IR system should go. As an example, he asked whether joins and set computations are really needed. In order to advance the field, he believes that what is needed is a common heterogeneous enterprise search collection consisting of an actual Intranet website, enterprise DBs, email, shared file systems contents, etc, together with actual queries. He hopes that such collection could be assembled and be used for making research progress, maybe in the context of a new TREC track.

XML and IR

The XML and IR morning session consisted of 4 papers that were selected by the workshop PC[1]. David Carmel from IBM Haifa Research Lab opened the session by presenting a paper co-authored with Broder, Maarek, Mandelbrod and Mass, which dealt with "*Extending the XML Fragment Model to Support Querying over Annotated Text*". The motivation for that work is to adapt and extend the XML Fragment model (introduced at a previous edition of the workshop and formalized at SIGIR'03) so as to accommodate for text annotations, which might be overlapping. Carmel first reminded to the audience the key ideas behind the XML Fragment model, where XML collections are queried via pieces of XML, or XML fragment, of the same nature as the documents forming the collection, in the pure IR spirit. He illustrated his purpose by giving a few examples of how information needs are expressed in XML Fragments. He then explained how certain automatic semantic annotators, such as developed in the context of IBM's Unstructured Information Management Architecture (UIMA) effort, do not generate valid XML. Indeed the annotations being generated on a same piece of text can easily be overlapping, and overlapping tags cannot be represented in XML either at the document or at the query levels. To answer this need, Carmel and his colleagues proposed to introduce new operators, namely for the intersection and concatenation of tags. He detailed the syntax and semantics of these operators and a small discussion started where several workshop participants expressed their interest and discussed how they had been confronted with similar constraints in computational linguistics applications.

The second presentation was given by Ludovic Denoyer, from University Paris 6, and dealt with "*Document Mapping for Heterogeneous Corpora*". This work was conducted jointly with Wisniewski and Gallinari at the LIP6 Lab. Denoyer first explained how dealing with heterogeneous XML corpora is pretty complex due to the variety of schemas or DTDs. In order to allow search engines to handle different formats or document types, he proposed as an option to use mediators, which would transform queries and documents. The ultimate goal is to find a relation between different structures and schemas, or in other terms, address the "schema matching task" (as known in the DB community) from an XML perspective. Denoyer explained that this task can be divided into the following sub-tasks: (1) learn the correspondences between tags, and (2) match tags in a "1-on-1" or "n-to-m" manner. However, in IR the context and tasks are different, indeed, the source DTD/schema is often unknown, the order of content elements might be meaningful, the number of tags might be much larger than in DBs (in the INEX corpus there are more than 100 tags), tag names often carry weak semantic, etc. Due to the high complexity, and the different needs, he argued that in most cases, approximate mapping may be sufficient, which thus allows for machine learning techniques. He proposes to learn the mapping from the data, and have it depend on both structure and content. In this approach, the document model is represented stochastically, via two probability distributions: a structural probability and a content probability. The structural probability is computed as a join probability (e.g., so as to model the probability to see a "title" tag under a "doc" tag), and a content probability, using a local naïve Bayes model. Then, in

[1] The XML and IR workshop PC consisted of M. Consens, M. Gertz, T. Grabs, D. Hawking, J. Jose, M. Lalmas, Y. Mass, P. Raghavan, and D. Suciu, whom we thank for their thorough work here.

order to find the structure of a new document, mediator tags that maximize the probability of a document need to be learned. He concluded showing the experiments that he and his colleagues run on the INEX corpus, by splitting it and removing part of the tags in order to simulate a heterogeneous corpus. The results were encouraging and demonstrated that using a combination of structure and content was best. He still believes though, that more tests and real heterogeneous corpora are needed before fully validating his approach.

The third paper of the session was presented by Karen Sauvagnat, from IRIT. This work, jointly authored with Mohand Boughanem, discussed *"The impact of leaf nodes relevance values evaluation in a propagation method for XML retrieval*".
Sauvagnat first reminded the fundamental difference between the DB and IR views in terms of the roles of relevance and efficiency in XML retrieval. She surveyed previous work by Goevert *et al* in 2002, on augmentation methods, where indexing weights are assigned to inner nodes by propagation from leaves. She introduced then the XFIRM model, where, in the same spirit, all leaf nodes are indexed, and the weight of inner nodes is computed based on the distance that separates nodes in the tree. Sauvagnat also described both, the logical and physical data representation models, used in her approach. The XFIRM query language is based on a complete query language (keywords with phrases, AND & OR operators, structural conditions, and even hierarchical conditions in its more complex form) that looks like a variant of XPath. She also detailed the query processing mechanism, where each query is decomposed in elementary sub-queries that are then processed independently. Each node is consequently assigned a relevant value, which is propagated up the tree, and then sub-queries are reconstructed (using the nearest common ancestor for instance). Sauvagnat also presented the experiments her team conducted on the INEX collection, with the strict CAS measure, which were positive. She concluded by explaining that this work should be considered as a starting point for finding an appropriate formula for leaf nodes relevance values evaluation. Possible directions of future research would be to evaluate different formulas taking into account the type of the elements to be returned, the length of the element, etc.

The final presentation of the session was given by Huyen-Trang Vu, from Paris 6 University. She discussed her joint work with Piwowarski and Gallinari, also from LIP6, entitled *"Filtering in XML Retrieval: a Prospective Analysis*". One key challenge of XML retrieval is to reduce or avoid overlapping results induced by nested elements. In this work, she proposes to use filtering as a post-processing mechanism to remove redundant results. The post-processing mechanism consists of three phases: re-scoring, filtering and final scoring. Re-scoring is based on the "utility theory" as defined by Crestani *et al*, and a novelty score to be balanced with the original relevance score. Filtering can be conducted in two manners: the first type is related to adaptive filtering, while the second is based on a recursive process of the document tree. Both strict and "tolerant" filtering techniques (where some overlaps are allowed) are considered. Finally, she discussed two different strategies for overall scoring. Vu then explained how she plans to use the INEX collection to analyze nesting and analyze and compare various filtering strategies.

IR and DB Workshop

The afternoon session was dedicated to the First WIRD workshop (co-chaired by Thomas Roelleke and Arjen P. de Vries) and consisted of three papers selected by the workshop PC[2]. Ingo Frommholz from Fraunhofer IPSI started by presenting COLLATE an indexing and retrieval system for an historical film archive. It provides film researchers with a collaborative environment in which historic documents about European films can be analyzed, interpreted, and discussed, using nested annotations and discourse structure relations among them. First, discussion about films is represented by free text annotations. Also, the system indexes the annotations of annotations, which is where the discourse structure enters the game. At the moment, the collection studied consists of almost 7000 documents and 2000 annotations. The COLLATE system has been implemented using four-valued probabilistic Datalog, which distinguishes itself from other approaches by basing the inference on the open-world assumption: negative and positive evidence are treated independently, so the absence of *p(a)* does not automatically imply *not(p(a))*. Ingo then explained how query processing takes place in the current prototype implementation. Further research is planned to validate experimentally the impact and significance of using annotations as proposed.

Vojkan Mihajlović, a PhD student at University of Twente, presented the second paper in the afternoon, posing the question whether an 'XML-IR-DB sandwich' would taste better with 'an algebra in between'. He argued that the traditional, layered database architecture is very suited for the development of XML-IR systems, and allows the development of structured document retrieval systems on top of relational database systems. The paper focuses specifically on the role of the logical level in the proposed system architecture. This intermediate layer between the XML level and the relational database system serves three goals: expressing XPath expressions and about clauses requires operators that are not available in the relational database system, it allows for query optimization, and, it allows expressing the propagation of scores through the query expression. He presented first a 'Boolean' region algebra to explain the general idea, and then moved on to a ranked region algebra for expressing more powerful retrieval models. Vojkan finally motivated the potential for query optimization based on some well-chosen examples of rewriting expressions using the algebraic properties of the operators.

[2] The WIRD'04 PC consisted of S. Chaudhuri, I. Frommholz, N. Fuhr, T. Grabs, D. Hiemstra, U. Kruschwitz, A. Natsev, J. Shanmugasundaram, D. Suciu, and M. Taylor.

The final talk of the afternoon, given by Cheng Xiang Zhai from the University of Illinois at Urbana-Champaign, motivated Entity Retrieval from a collection consisting of both structured and unstructured data. The specific task studied in the paper is to find all web pages for a certain researcher (the entity). When compared to 'traditional' database search, the entity retrieval problem is challenging because it would require search at a semantic level instead of the syntactic one that can be expressed in SQL queries, e.g. due to ambiguous names. When compared to full-text retrieval, the difference is that entity retrieval involves a special type of information need, more objectively defined than the one usually studied in information retrieval experiments. Cheng Xiang and his students have shown in their research work that a full-text search in a (newly constructed) test collection of web pages improves when taking structured information from the DBLP data into account for query expansion.

Discussion

The joint workshop ended with an open discussion among participants, with the main goal to identify the issues in the integration of databases and information retrieval, and the potential role of XML. We opened the discussion by asking why the workshops had received a smaller number of submissions than expected (8 for the XML and IR workshop and 4 for WIRD). It was not clear whether it was due to the large number of workshops held at SIGIR or the fact that, while sufficient interest exists in the main topics addressed by the workshops, the area of interest might better be captured under the notion 'enterprise search'. Here, David Hawking advertised the upcoming Enterprise Search task at TREC, intended to (eventually) replace the web-track. He gave some insight in the plans, notably the data most likely to be used (W3C collection, and maybe cleaned-up Enron data), and the track set up (includes a collaborative relevance assessment phase).

We proceeded by eliciting the audience's opinion on the most important topics that require further research. This session lead to a varied number of issues, summarized here into the following partial list:

- Semi-structured search, 'intelligent search', 'enterprise search'.
- What do the appropriate query languages look like?
- What is the amount of heterogeneity (in structure) of the data to be searched?
- How do we evaluate proposed techniques?
- How to achieve efficient and scalable search systems?
- Do we need ontologies?
- How does this all relate to 'customer relationship management'?

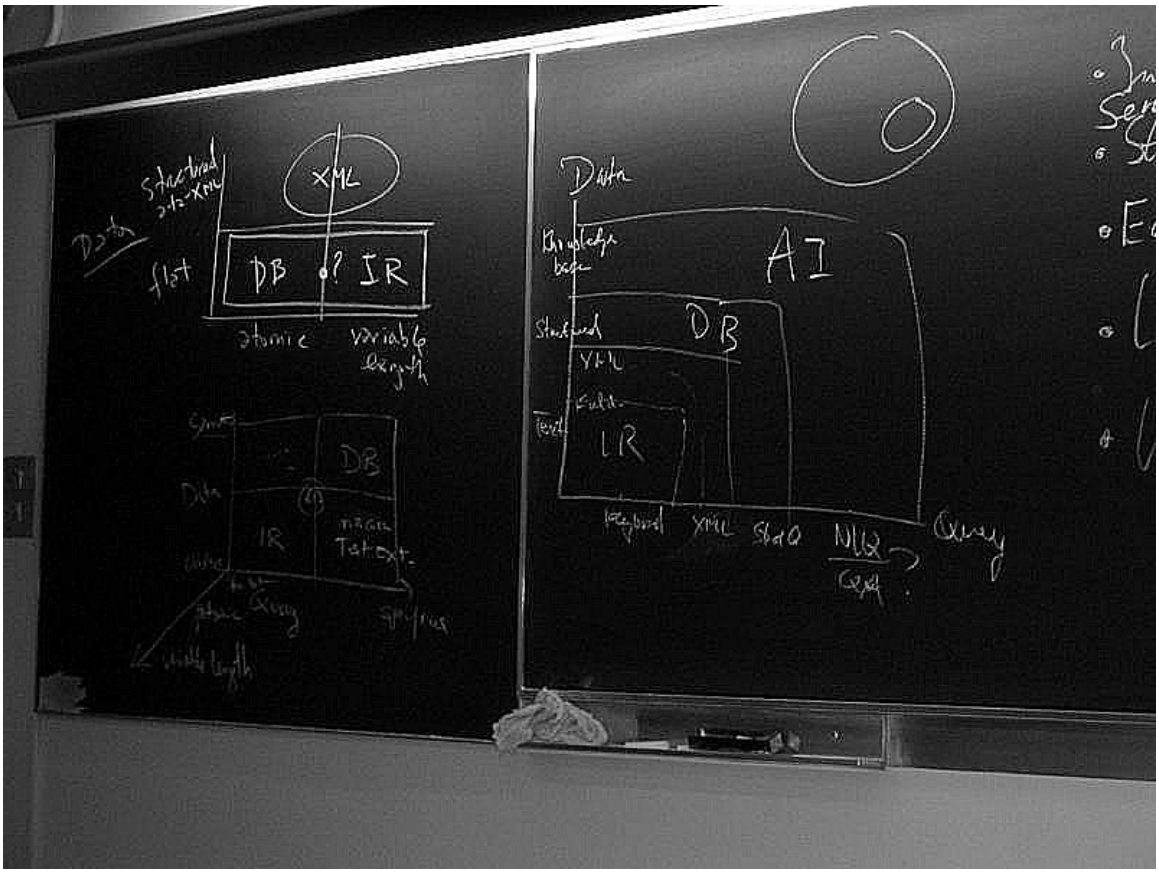

Figure 1: Snapshot of the board during the discussion

Next, a decentralized brainstorm session started in an attempt to classify the wide variety of issues found into a more organized structure. At some time during this discussion, the graphics depicted above 'appeared' on the blackboard; notably a graph showing 'expressiveness of the query' on the X-axis ('keywords', 'XPath (+about)', 'structured queries (+about)' and eventually 'natural language') and 'amount of structure in the data' on the Y-axis ('text', 'XML', 'structured', 'knowledge base'), where IR would cover the lower left corner, then XML-IR a wider area, DB an even wider area, and artificial intelligence being the field that covers all topics… although not all of us agree on this latter view!

Summarizing the opinions expressed in the group discussion, most workshop participants seem to view the 'integration of DB and IR' as an intermediate step toward truly 'intelligent search', and this intermediate step being the missing link to build search engines for enterprise data.