08111 Abstracts Collection
Ranked XML Querying
— Dagstuhl Seminar —

Sihem Amer-Yahia[1], Divesh Srivastava[2] and Gerhard Weikum[3]

[1] Yahoo Research New York, US
[2] AT&T Florham Park, US
divesh@research.att.com
[3] MPI Saarbrücken, DE
weikum@mpi-sb.mpg.de

Abstract. From 09.03. to 14.03.08, the Dagstuhl Seminar 08111 "Ranked XML Querying" was held in the International Conference and Research Center (IBFI), Schloss Dagstuhl. During the seminar, several participants presented their current research, and ongoing work and open problems were discussed. Abstracts of the presentations given during the seminar as well as abstracts of seminar results and ideas are put together in this paper. The first section describes the seminar topics and goals in general. Links to extended abstracts or full papers are provided, if available.

Keywords. Scoring methods for XML, Ranking approximate XML answers, Top-K query processing, Querying structured and unstructured data, XML Full-Text Querying, Querying heterogeneous XML, Extracting structure from unstructured data, Text mining, XML data integration

08111 Report – Ranked XML Querying

This paper is based on a five-day workshop on "Ranked XML Querying" that took place in Schloss Dagstuhl in Germany in March 2008 and was attended by 27 people from three different research communities: database systems (DB), information retrieval (IR), and Web. The seminar title was interpreted in an IR-style "andish" sense (it covered also subsets of Ranking, XML, Querying, with larger sets being favored) rather than the DB-style strictly conjunctive manner. So in essence, the seminar really addressed the integration of DB and IR technologies with Web 2.0 being an important target area.

Keywords: Scoring methods for XML, Ranking approximate XML answers, Top-K query processing, Querying structured and unstructured data, XML Full-Text Querying, Querying heterogeneous XML, Extracting structure from unstructured data, Text mining, XML data integration

Dagstuhl Seminar Proceedings 08111
Ranked XML Querying
http://drops.dagstuhl.de/opus/volltexte/2008/1536

Joint work of: Amer-Yahia, Sihem; Hiemstra, Djoerd; Roelleke, Thomas; Srivastava, Divesh; Weikum, Gerhard

Full Paper: http://drops.dagstuhl.de/opus/volltexte/2008/1535

08111 Summary of the Breakout Session on Ranking, Scoring and Real-life Applications

Here are the minutes of the session on Ranking vs scoring and real life applications

They are as typed in the meeting so they might need some cleaning.. The summary topics and the their interaction are at the end of the document

Thanks

-ihab

Harold * the need for a good ranking algorithms for XML that can be generally applied in a product and takes into account the structure (not contents only)

- same thing exists in treating relational DB as graphs (keyword search -> steiner trees).
- ranking is more important than scoring

Djeord * Entity retrieval vs. scoring the occurrence of a specific element

- Ranking a "person" in an expert search
- Extraction of "entities" might involve uncertainty
- Scores and metrics reflect confidence that "it is really an entity" + evidence of expertise

Thomas * Ranking can be interpreted as uniform while it is not.. scores add more semantics (how far the second from the first)

- Meaningfulness of ranking w.r.t associated scores
- Algebraic description modeling of ranking strategies -> probabilistic scoring + Ranking as input to another process
- Interpreting scores/ relevance as evidence of ranking e.g. viewing it as voting

Kostas * Add scoring to ranking with preferences (how much I prefer red to green)

Ihab * people do not think in terms of total orders but would like to see the results that way

- mismatch between user-feedback and preference specs and processing of total orders
- Need natural specification (preference languages is a step there.. but lead to inefficient algorithms)
- meaningfulness of rank-agregation functions (score aggregation?)

Emiran * what applications in which a scoring function is useful in that particular application, which scoring function should I use

- plugging in black-box scoring function
- declaring properties of scoring functions/process so they can be matched against application needs

Divesh * in IR, language models exists that can back up ranking functions.. how about XML? defining the notion of a meaningful answers -> snippets vs. whole documents granularity of the results -> sites do group documents from the same site for presentation Holistic Ranking e.g., ranking with diversity or top 10 centroids (10 good answers collectively) vs. (10 best answers in a race)

Maarten and Martin*

notion of an answer (meaningful answer) entry points

Topics:

(I) Ranking Functions Properties and Meaningful scores

- Natural ranking functions from user expression to total order production
- matching properties to applications
- meaningfulness of raked results (top 10 with 9 irrelevant results)

Query processing of Ranking

- effect of ranking function on devising plans and vice versa
- properties of ranking functions

(II) Evidence-based ranking and aggregation

- Entity search
- aggregating structure and contents ranking in XML
- Probabilistic Ranking

(III) Ranking granularity

- Entry points
- representation of answers

Holistic Ranking

- what construct a good top-k objects
- Capturing ranking and diversity of results

-

aggregation complexity and meaningfulness +——————————> Evidence | | | Mutual effect between holistic | ranking functions and processing goals Independence | ranking that changes the answer sets (approximation trading efficiency to quality) and structure | complexity|| | rank-aware Query Processing | V Answers

Joint work of: Ilyas, Ihab

4.5min of belief and dream and... 30sec of realism - Making DB and IR (socially) meaningful

Sihem Amer-Yahia (Yahoo! Research - New York, US)

A brief overview of my current work on social recommender systems

Keywords: Recommender systems, ranking, social

Why should (XML Retrieval) Effectiveness Measures Matter?

Mariano P. Consens (University of Toronto, CA)

Evaluation measures are used to evaluate the effectiveness of IR systems and try to answer the question of how relevant are the answers produced by systems to the users information needs. Effectiveness measures are applied in the context of an evaluation methodology that includes a data collection, a specific task, relevance assessments (by users or experts) and assumptions about user behaviour.

In this talk we motivate and discuss our recent work in developing evaluation measures for XML Retrieval that can adapt to a variety of structures collections and tasks, and can also accommodate different models of user behaviour.

XML Distributed Retrieval

Emiran Curtmola (University of California - San Diego, US)

As the web evolves, it is becoming easier to form communities based on shared interests, and to create and publish data on a wide variety of topics.

With this "democratization of information creation" comes the natural desire to make one's data accessible for querying within the community and also be able to query the XML global collection that, virtually, is the union of all local data collections of others within the community.

We propose a distributed infrastructure in which data resides only with the publishers owning it who thereby maintain control on who can access and how to advertise their data. The infrastructure disseminates user queries to publishers, who answer them at their own discretion.

Given the virtual nature of the global data collection, we study the challenging problem of efficiently locating publishers in the community that contain data items matching a specified query. We propose a distributed index structure, UQDT, that is organized as a union of Query Dissemination Trees (QDTs), and realized on an overlay (i.e., logical) network infrastructure. Each QDT has data publishers as its leaf nodes, and overlay network nodes as its internal nodes;

each internal node routes queries to publishers, based on a summary of the data advertised by publishers in its subtrees.

We experimentally evaluate design tradeoffs, and demonstrate that UQDT can maximize throughput by preventing any overlay network node from becoming a bottleneck. We conclude by posing the problem of XML distributed ranked retrieval to discover only the topK relevant publishers, and moreover, the topK matching documents.

Keywords: XML community distributed search P2P peer-to-peer

Efficient graph-based context-sensitive search

Debora Donato (Yahoo Research - Barcelona, ES)

We study the problem of context-sensitive search, in which a query consists of a set of terms q and a document p already in the collection, and the task is to rank the documents in decreasing order of relevance with respect to the query hq, pi.

The document p provides the context in which the query terms q should be understood. By attaching a context to the query terms, the search results of a search initiated in a particular page can be made more relevant.

If we consider the collection of documents as a directed graph G with vertices the documents and edges the links among the documents, an intuitive approach for the problem is to employ the personalized PageRank computation by defining the teleport vector to be the vector that jumps back to the query document p. However, computing and storing personalized PageRank vectors for each documents separately is infeasible. Instead we compute a partition of G by using a spectral method and we store a personalization vector only for each cluster. In addition to these approximate personalization vectors we use also traditional informationretrieval measures such as BM25 and other graph-based similarity measures to form feature vectors. RankSVM is then used to learn weights for individual features given suitably constructed training data. Documents are ranked at query time using the inner product of the feature and the weight vectors.

The experiments made using Wikipedia indicate that the proposed method considerably improves results obtained by a more traditional approach that does not take the context into account.

Probabilistic, Object-oriented Logics for Annotation-based Structured Document Retrieval and Discussion Search - the POLAR Framework

Ingo Frommholz (Universität Duisburg-Essen, DE)

In this talk I introduce POLAR, a probabilistic object-oriented logical framework for annotation-based information retrieval.

POLAR allows for modelling so-called structured annotation hypertexts, which regard documents and annotations as interlinked structured documents (which are possibly described in XML). POLAR supports querying structured annotation hypertexts as well as annotation-based information retrieval applying probabilistic inference and knowledge and relevance augmentation. POLAR's semantic is based on Modal logics using an extension of Kripke structures, and it is implemented upon four-valued probabilistic Datalog. Experimental results show the effectiveness of the built-in annotation-based document and discussion search approaches.

Keywords: POLAR, annotation-based retrieval, four-valued probabilistic datalog, modal logics

Personalizing XML Full Text Search in PIMENTO

Irini Fundulaki (FORTH - Heraklion, GR)

In PIMENTO we advocate a novel approach to XML search that leverages user information to return more relevant query answers. This approach is based on formalizing *user profiles* in terms of *scoping rules* which are used to rewrite an input query, and of *ordering rules* which are combined with query scoring to customize the ranking of query answers to specific users.

Keywords: XML Full Text Search, Personalization

Joint work of: Fundulaki, Irini; Amer-Yahia, Sihem; Laks, Lakshmanan

Extended Abstract: http://drops.dagstuhl.de/opus/volltexte/2008/1534

Full Paper:
http://ieeexplore.ieee.org/search/wrapper.jsp?arnumber=4221739

See also: @InProceedingss.07:_person_xml_full_text_searc_in_pimen,author=S. Amer-Yahia, I. Fundulaki and L. Lakshmanan,title =

Sound ranking algorithms for XML search in PF/Tijah

Djoerd Hiemstra (University of Twente, NL)

When designing ranking algorithms for PF/Tijah (PF/Tijah is a text search extension for MonetDB/XQuery developed at the University of Twente) it turned out to be difficult to come up with approaches to ranking that reflect the actual combined content and structure constraints of queries, while at the same time producing equal rankings for queries that are semantically equal. Ranking algorithms that produce different rankings for queries that are semantically equal are easily detected by tests on large databases: We call such algorithms *not sound*.

We report the behaviour of 120 different approaches to ranking content-and-structure queries on 7 pairs of queries for which we expect equal ranking results. We show that 111 out of these 120 approaches are not sound, i.e., they fail to produce equal rankings in the cases studied. Of the remaining 9 approaches, only 2 adhere to the W3C XQuery Full-Text standard, which requires matching semantics, and which requires retrieval scores to be smaller or equal than 1 at all times. The difficulties in implementing effective and sound ranking for XQuery Full-Text might affect its acceptance as a standard in the future.

Keywords: XML Information Retrieval, XQuery Full-Text

Joint work of: Hiemstra, Djoerd; Klinger, Stefan; Rode, Henning; Flokstra, Jan; Apers, Peter

Sound ranking algorithms for XML search in PF/Tijah

Djoerd Hiemstra (University of Twente, NL)

We argue that ranking algorithms for XML should reflect the actual combined content and structure constraints of queries, while at the same time producing equal rankings for queries that are semantically equal. Ranking algorithms that produce different rankings for queries that are semantically equal are easily detected by tests on large databases: We call such algorithms *not sound*. We report the behaviour of different approaches to ranking content-and-structure queries on pairs of queries for which we expect equal ranking results from the query semantics. We show that most of these approaches are not sound. Of the remaining approaches, only 3 adhere to the W3C XQuery Full-Text standard.

Keywords: XML Information Retrieval, XQuery Full-Text

Joint work of: Hiemstra, Djoerd; Klinger, Stefan; Rode, Henning; Flokstra, Jan; Apers, Peter

Full Paper: http://drops.dagstuhl.de/opus/volltexte/2008/1533

URank: Ranking and Aggregation Queries in Probabilistic Databases

Ihab Ilyas (University of Waterloo, CA)

Ranking and aggregation queries are widely exploited in data exploration, data analysis and decision making scenarios. While most of the currently proposed ranking and aggregation techniques focus on deterministic data, several emerging applications involve data that is unclean or uncertain. Ranking and aggregating uncertain (probabilistic) data raises new challenges with respect to query semantics and processing, which makes conventional methods inapplicable.

In this talk, I will introduce new formulations for ranking and aggregation queries in probabilistic databases. The new formulations are based on marriage of traditional ranking and aggregation algorithms with possible worlds semantics. In the light of these formulations, I will describe a generic processing framework supporting both query types, and leveraging existing query processing and indexing capabilities in current database systems. The framework encapsulates a state space model, and efficient search algorithms that compute query answers with optimality guarantees.

Keywords: Ranking, top-k, probabilistic, uncertain

See also: Ihab Ilyas is an Assistant Professor of Computer Science at the University of Waterloo since July 2004. He received his PhD in computer science from Purdue University, West Lafayette in 2004. He holds BS and MS degrees in computer science from Alexandria University, Egypt. He spent two summers with IBM Almaden Research Center and he is currently an IBM CAS faculty fellow since January 2006. His main research is in the area of database systems, with special interest in top-k and rank-aware query processing, managing uncertain and probabilistic databases, self-managing databases, indexing techniques, and spatial databases. For more information and a list of publications, please visit Ihab's web page.

Structuring and Ranking Opinions using Econometrics

Panos Ipeirotis (New York University, US)

Today, users post online reviews expressing their opinions for movies, restaurants, and many other products. They also evaluate merchants and react to news about political campaigns.

Structuring and ranking these opinions in terms of importance and polarity is a difficult research problem.

How can we infer the importance and polarity of the posted content? How can we structure and quantify the effect of the online opinions?

Many existing approaches rely on human annotators to evaluate the polarity and strength of the opinions, a laborious and error-prone task. We take a different approach by considering the economic context in which an opinion is evaluated. We rely on the fact that the text in on-line systems influence the behavior of the readers and this effect can be observed using some easy-to-measure economic variables, such as revenues or product prices.

Then, by reversing the logic, we infer the semantic orientation and the strength of an opinion by tracing the changes in the associated economic variable. In effect, we combine econometrics with text mining algorithms to identify the "economic value of text" and assign a "dollar value" to each opinion, quantifying sentiment effectively and without the need for manual effort.

We make the discussion concrete by presenting results on reputation systems, on product reviews, and on how political campaigns are affected by online chatter in blogs and mass media.

Keywords: Opinion ranking, econometrics

Keyword Proximity Search over Data Graphs

Benny Kimelfeld (The Hebrew University of Jerusalem, IL)

Various systems support keyword search over different types of data with some degree of structure, e.g., relational databases and XML graphs. The quality of search results is determined by the frequency of the keywords as well as the associations among the keywords in the (structured) result. The number of elements that participate in the association, as well as the nature of the links that interconnect the elements, are major factors in the strength (relevancy) of a result.

In this talk, I will describe an abstract framework that provides a formal setting for analyzing the performance and quality of search engines. Within that framework, I will survey various algorithmic approaches and complexity results regarding the implementation of engines. Finally, I will discuss the future challenges.

Extending an XQuery compiler with Full-Text capabilities

Stefan Klinger (Universität Konstanz, DE)

The Pathfinder project aims at the creation of a purely relational XQuery processor, to this end translating the XML data into relations, and XQuery queries into relational algebra plans. Current results show that this approach can indeed lead to a highly scalable, high performance XQuery database.

In this talk I'm going to present the $Pathfinder^{FT}$ project, the goal of which is to extend the Pathfinder XQuery compiler with an infrastructure suitable for dealing with XQuery Full-Text.

Instead of providing one particular XQuery Full-Text implementation that fixes the involved information retrieval and scoring models, the implementation defined concepts shall be left as gaps in a more versatile framework.

The current $Pathfinder^{FT}$ prototype implements a complete infrastructure for IR and scoring models, i.e., it allows for propagation of scores along XPath axis steps, their migration from predicates, and their combination when used with XQuery operators like Boolean or set operations.

Keywords: RDBMS XQuery Full-Text

Multi-Dimensional Search for Personal Information Systems

Amélie Marian (Rutgers Univ. - Piscataway, US)

With the explosion in the amount of semi-structured data users access and store in personal information management systems, there is a need for complex search tools to retrieve often very heterogeneous data in a simple and efficient way. Existing tools usually index text content, allowing for some IR-style ranking on the textual part of the query, but only consider structure (e.g., file directory) and metadata (e.g., date, file type) as filtering conditions. We propose a novel multi-dimensional approach to semi-structured data searches in personal information management systems by allowing users to provide fuzzy structure and metadata conditions in addition to keyword conditions. Our techniques provide a complex query interface that is more comprehensive than content-only searches as it considers three query dimensions (content, structure, metadata) in the search. We propose techniques to individually score each dimension, as well as a framework to integrate the three dimension scores into a meaningful unified score. Our work is integrated in Wayfinder, an existing fully-functioning file system.

In this talk, I will present our scoring framework and query processing strategies and discuss the effect of approximating individual dimensions on the overall scores and ranks of files, as well as on query performance.

An Adaptive XML Retrieval System

Yosi Mass (IBM - Haifa, IL)

The eXtensible Markup Language (XML) can be used to add structure to full-text documents in the form of hierarchical tagging. From an information retrieval perspective, the challenge in XML retrieval is to exploit the induced structure to return the most relevant parts inside documents that best fit user needs. Several XML retrieval methods have been proposed, ranging from indexing all XML elements into a single index to approaches that index only leaf elements and propagate their scores to parent elements. We take a different approach and suggest an adaptive indexing schema that uses existing IR engines as a meta-search engine to achieve an XML retrieval system. This solution is easy to deploy and has a solid theoretical background based on existing research of classical IR for full-text documents. We describe the adaptive XML retrieval system and discuss the results of experiments done on several XML benchmark collections

Keywords: XML, IR, Ranking, Adaptive, Scalability

Joint work of: Mass, Yosi; Shmueli-Scheuer, Michal

Modelling XML Retrieval in Probabilistic Logical Abstraction Layers

Thomas Rölleke (Queen Mary College - London, GB)

XML retrieval requires to model content and structure, and XML derivates such as RDF require to model semantics. The expressiveness of relational and object-oriented data models supports the modelling of XML retrieval, and probabilistic facilities help to model retrieval models (ranking functions). Therefore, XML retrieval is a motivation for research on probabilistic extensions of abstract data models such as the relational algebra, SQL, Datalog and object-oriented logics. These probabilistic logical abstraction layers should support both, the modelling of retrieval models (TF-IDF, probabilistic model, BM25, language modelling), and the modelling of retrieval tasks (XML retrieval, classification, summarisation). The talk will address aspects of probability aggregation and estimation, and the application of probabilistic extensions of SQL and probabilistic relational algebra will be demonstrated.

Relevance Feedback in the TopX Search Engine

Ralf Schenkel (MPI für Informatik - Saarbrücken, DE)

Keyword-based queries are an important means to retrieve information from XML collections with unknown or complex schemas. Relevance Feedback integrates relevance information provided by a user to enhance retrieval quality. For keyword-based XML queries, feedback engines usually generate an expanded keyword query from the content of elements marked as relevant or nonrelevant. This approach that is inspired by text-based IR completely ignores the semistructured nature of XML. This talk presents a framework that expands a keyword query into a full-fledged content-and-structure query, making the important step from pure content-based to structural feedback. It gives an overview of experiments done with the established INEX benchmark and our TopX search engine.

Joint work of: Schenkel, Ralf; Theobald, Martin; Sammodi, Osama

What can an XML database do about ranked XML retrieval?

Harald Schöning (Software AG - Darmstadt, DE)

Being an XML database, Tamino does boolean retrieval unlike typical retrieval systems. Nevertheless, ranking is an important functionality for our customers.

In my talk I will sketch the potentials and limitations we found for statisfactory ranked XML retrieval

Using Domain Knowledge to Extract Wrappers for Tree-Structured Documents

Pierre Senellart (Telecom Paris Tech, FR)

We present an original approach to automatic, unsupervised, wrapper induction for tree-structured documents (e.g., HTML result pages for services of the deep Web). A domain knowledge for the specific domain of interest is assumed, in a specific format, easy to obtain for standard applications. This domain knowledge is used as the input of a gazetteer that linearly annotates parts of the document with domain concepts. This annotation, both imprecise and imperfect, is then fed to an unsupervised machine learning algorithm (conditional random fields for XML) that uses the tree structure of the document, and its repetitive parts, to correct and generalize this annotation. Experiments show the viability and potential of the approach.

Keywords: Information extraction, deep Web, machine learning

Joint work of: Senellart, Pierre; Muschick, Daniel; Gilleron, Rémi; Tommasi, Marc

Multidimensional Contextual Preferences

Kostas Stefanidis (University of Ioannina, GR)

To provide users only with relevant data from the huge amount of available information, personalization systems utilize preferences to allow users to express their interest on specific pieces of data. Most often, user preferences vary depending on the circumstances. For instance, when with friends, users may like to watch thrillers, whereas, when with their kids, they may prefer to watch cartoons. Contextual preference systems address this challenge by supporting preferences that depend on the values of contextual attributes such as the surrounding environment, time or location. In this talk, I will present our model for expressing such contextual preferences. We model context using a set of hierarchical attributes, thus allowing the expression of context information with various levels of detail. Then, we define the context resolution problem as the problem of (a) identifying those preferences that qualify to encompass the context of a query and (b) selecting the most appropriate among them. We propose an algorithm for context resolution based on a data structure, called profile tree, that indexes preferences according to their associated context.

I shall also touch upon our current work on efficient scoring of database tuples based on contextual preferences, which is based on selecting and pre-computing representative rankings. In particular, we exploit the hierarchical nature of context attributes to identify representative contexts.

Furthermore, we introduce a method for grouping preferences based on the similarity of the scores that they produce. This method uses a bitmap representation of preferences and scores with various levels of precision that lead to approximate rankings with different degrees of accuracy.

TopX 2.0 - A (Very) Fast Object-Store for Top-k XPath Query Processing

Martin Theobald (Stanford University, US)

TopX is an efficient top-k engine for XPath-like full-text queries. It supports a probabilistic-IR scoring model over full-text content conditions with precomputed relevance weights for combined tag-term pairs, path conditions for all XPath axes as exact or relaxable constraints, and ontology-based relaxation of tags and terms as similarity conditions for ranked retrieval. For speeding up top-k queries, various techniques are employed: probabilistic models as efficient score predictors for a variant of Fagin's threshold algorithm, cost-based scheduling of sequential and random accesses to disk-resident inverted index structures, as well as incremental merging of index lists for dynamic and self-tuning query expansion.

While the original TopX prototype has been focusing on top-k query evaluations on top of a relational backend, TopX 2.0 introduces a highly optimized, object-oriented storage for XML with direct access to customized inverted files. The talk presents our brand-new, multiple nested block-index structure which seamlessly integrates top-k-style sorted access to large disk blocks with in-memory merge-joins for efficient score aggregations.

DB&IR from a DB Viewpoint

Gerhard Weikum (MPI für Informatik - Saarbrücken, DE)

This talk gives a subjective overview of recent research on the integration of database system (DB) and informmation retrieval (IR) methods, and explains why this requires introducing ranking into the DB world.

From a DB viewpoint, ranking is desirable for four major reasons. First, when adding text-matching conditions to structured query languages like SQL or XPath, similarity scoring and result ranking are needed based on keyword frequency statistics. Second, when exploring structured databases by personal preference search, top-k querying is highly desired to avoid returning too many answers. Third, when querying multiple heterogeneous data sources, the lack of a unified global schema entails the need for schema-free or schema-relaxation-tolerant search with approximate results. Fourth, when performing information extraction on semistructured or natural-language text sources in order to facilitate entity search, ranking needs to consider confidence measures of extracted facts, statistics over entity-relation graphs, and the compactness of subgraphs connecting related entities.

Each of these directions has led to promising results and is driving ongoing research work, selectively surveyed by the talk.

A Parameterised Search System

Arjen P. de Vries (CWI - Amsterdam, NL)

The talk introduces the concept of a Parameterised Search System (PSS), which allows flexibility in user queries, and, more importantly, allows system engineers to easily define customised search strategies. Putting this idea into practise requires a carefully designed system architecture that supports a declarative abstraction language for the specification of search strategies. These specifications should stay as close as possible to the problem definition (i.e., the retrieval model to be used in the search application), abstracting away the details of the physical organisation of data and content. We show how extending an existing XML retrieval system with an abstraction mechanism based on array databases meets this requirement.

Full Paper:
http://www.springerlink.com/content/6054w43473r12h07/