Report on the DB/IR Panel at SIGMOD 2005

Sihem Amer-Yahia
Moderator
AT&T Labs Research, USA

Pat Case
Library of Congress, USA
Thomas Rölleke
Queens Mary University, UK
Jayavel Shanmugasundaram
Cornell University, USA
Gerhard Weikum
Max Plank Institute, Germany

1. MOTIVATION

This paper summarizes the salient aspects of the SIGMOD 2005 panel on "Databases and Information Retrieval: Rethinking the Great Divide". The goal of the panel was to discuss whether we should rethink data management systems architectures to truly merge Database (DB) and Information Retrieval (IR) technologies. The panel had very high attendance and generated lively discussions.[1]

Until now, the DB and IR communities, while each very successful, have evolved largely independently of each other. The DB community has mostly focused on highly structured data, and has developed sophisticated techniques for efficiently processing complex and precise queries over this data. In contrast, the IR community has focused on searching unstructured data, and has developed various techniques for ranking query results and evaluating their effectiveness. Consequently, there has been no single unified system model for managing both structured and unstructured data, and processing both precise and ranked queries. Most prior integration attempts have "glued" together DB and IR engines without making fundamental changes to either engine.

However, emerging applications such as content management and XML data management, which have an abundant mix of structured and unstructured data, require us to rethink data management assumptions such as the strict dichotomy between accessing content in DB and IR systems. In fact, recent trends in DB and IR research demonstrate a growing interest in adopting IR techniques in DBs and vice versa. The goal of this report is to issue new challenges to both communities, in particular, from an application, end-user, querying and system architecture perspectives.

2. PANEL OVERVIEW

The panel included established DB and IR experts. We first list the set of questions asked to the panelists. We then present the viewpoint of each panelist and a summary of the discussion.

2.1 Panel Questions

1) Which real-world applications require a tight DB-IR integration? Can most applications be addressed by storing unstructured data as uninterpreted columns in a relational DB system, and invoking an IR engine over unstructured data?

2) XML is being touted as the dominant and pervasive standard that integrates structured and unstructured data, and XML query languages such as XQuery Full-Text [59], attempt to support this. Can we still cobble together a solution using traditional DB and IR systems? Or do we need to rethink the fundamental data management system architecture?

3) Does it make sense to evaluate "imprecise" queries over structured data and produce ranked results? Conversely, does it make sense to evaluate "precise and complex" queries over unstructured or semi-structured data? If so, do any of the IR techniques carry over to the structured domain, and vice versa? Does this then argue for or against a unified query model?

4) DB and IR systems are already complex pieces of software with decades of research and a strong commercial backing. Is it possible to design a clean underlying formal model (akin to the relational model and IR ranking models) that captures the whole gamut of issues that both classes of systems deal with? Is it feasible to build a system based on what could be exceedingly complex data and query models? Would this gain acceptance in the marketplace and displace loosely coupled DB and IR systems?

5) Are there any "cultural" issues that would prevent a true DB-IR unification?

2.2 Panel Discussion

The panelists selection covered different perspectives of the panel topic. Pat Case, a librarian at the Congressional Research Service at the U.S. Library of Congress, gave her expert user's view on combining full-text search with structured search. Then, Gerhard Weikum, a research director at the Max-Planck Institute of Computer Science in Saarbruecken, Germany, presented an applications' perspective on the integration of DB and IR technologies. The following panelist, Thomas Rölleke, a research fellow and lecturer at Queen Mary University in London, provided an IR-expert view on the panel topic. Finally, Jayavel Shanmugasundaram, an assistant professor at the Department of Computer Science at Cornell University, described his system architecture's viewpoints.

Pat Case, the first panelist, motivated the need for a search system that integrates DB and IR querying capabilities. She stated the fact that existing solutions lack some fundamental features needed by expert users who need to search a database of documents, such as the document repository at the Library of Congress, as opposed to searching the open Web. The first requirement of a good system is the ability to return fewer results since end-users must be

[1] Panel slides available at:
www.research.att.com/sihem/SIGMOD-PANEL/.

able to review all of the results. A good search system must allow users to refine their search results by explicitly limiting or expanding the number of answers or by using taxonomies and ontologies. The second requirement is the ability to parameterize the scoring method used to rank query answers. Most IR engines [2] are treated as black boxes which use proprietary scoring algorithms to decide, on behalf of end-users, how to rank query results. Pat argued for relevance that is based on user-specified criteria, not on some word frequency method such as tf*idf. As an example, a congressional bill is more relevant if it is of a certain bill type, if it has been reported out of committee, placed on calendar, discussed on the floor, passed by one chamber, has become law, has a large number of co-sponsors etc. In addition, a system should also permit exact and unscored searches. The third requirement is the need for ordered and unordered word distance operators. With the advanced search functionality provided in today's search engines, users get OR (which is useless, except for strings of synonyms, AND which is close to useless, NOT, which is dangerous and, PHRASE which is way too limiting and it is a lie in some systems! More generally, a search system should offer a full array of full-text search functionalities. In January 2005, the PEW Internet & American Life Project released a report titled: "Internet Searchers are confident, satisfied and trusting, but they are also unaware and naive". It noted that only 7% of users use more than 3 search engines on a regular basis. However, these are librarians, researchers, doctors, lawyers, scientists, academics, and the graduate students who need to know everything that has been written on their dissertation topic. Example functionality that would help such users is prefix, infix, and suffix wild cards, ordered and unordered distance operators, thesaurus integration, starts-with functionality, a usable NOT, and end user control over diacritics, case, and stop words. In addition to powerful text search primitives, Pat argued for the necessity to combine them with a full array of SQL-like searches on dates, numbers, strings, and nodes. Examples of such queries are date and number range searching and the ability to search within a single instance of a field or element. Right know, librarians are forced to choose between full-text and SQL-like search functionalities. At the Library of Congress (LoC), document metadata is ported from a relational database to a full-text search engine. As a result, the SQL search capabilities are lost. Finally, Pat argued for a standard end user syntax that combines structured and unstructured search and that can be used reliably across search systems.

The next panelist, Gerhard Weikum, described a number of applications such as customer support and health care management. In both cases, text such as problem descriptions (in customer support) or symptoms (in health care management) are connected to structured data such as location and time. Such applications thus require queries on both text and data. Moreover, they usually require ranked result lists rather than result sets. So the IR paradigm of ranked retrieval, based on probabilistic models of relevance, should be carried over to the world of structured data, too, and further lead to a unified ranking methodology for all kinds of combined information. This becomes even more important in the context of data integration. These days many scientific and business applications need to combine and analyze data that comes from different sources. Ideally, this would require reconciling schemas, identifying and linking matching entities in the data instances, and cleaning and transforming values. However, this kind of data integration is almost always the bottleneck, and often users would be gladly willing to work with less perfect data, with statistically "guessed" or "learned" matchings and approximately cleaned-up data values. An example of a technique that helps integration but naturally introduces such uncertainty is entity recognition which combines natural language processing methods with pattern matching and Markov-model-based learning in order to extract persons, products, etc. from text. Global queries on data sources that are partially and approximately integrated using such statistical and heuristic techniques naturally require ranked retrieval. Gerhard finished his presentation with a number of recommendations including (i) work on Approximate Query Processing, Statistics-based Information Extraction, (ii) integrate logic-based and statistics-based paradigms and establish foundations for probabilistic SQL and XQuery, (iii) develop system architectures for flexible scoring and ranking, (iv) develop cognitive models of user intentions and behavior, (v) develop a better experimental methodology towards reproducible results and more objective insights into efficiency/quality tradeoffs and, (vi) think about an integrated DB&IR curriculum. Finally, in order to address the "cultural" barrier Gerhard suggested to colocate the SIGMOD and SIGIR conferences.

The next panelist, Thomas Rölleke, argued that while DB research focuses on relational data modeling, SQL and transaction-based processing, IR research focuses on text document retrieval. As a result, although new trends such as multimedia applications and querying XML document collections are a driving force for the integration of IR and DB approaches, integrating both technologies in the same system is not feasible. This is also due to the fact that technology used in today's IT environments comprises vertical solutions for DB, enterprise, web and document search rather than integrated technology. IR yields the methods for relevance-based ranking, while DB research provides methods for dealing with structured, and, increasingly, semi-structured data. The integration is technologically challenging, and the question is whether an IR application on top of classical SQL technology meets the requirements and scalability of IR applications. Thomas argues that changes in the relational algebra core (management of uncertainty, stream-based processing) are needed for meeting IR requirements. Also, the cultural integration of the research communities is actually even more challenging than the technological integration. Thomas was one of the organizers of a SIGIR 2004 workshop on integration of DB and IR. However, he believes that while a unified DB and IR system is needed to improve expressiveness, scalability and abstraction, and, overall, productivity [20], as far as XML applications are concerned, XML on top of new relational IR technology works fine in practice.

The last panelist, Jayavel Shanmugasundaram, presented three alternative approaches for unifying DB and IR and argued that the first two options do not work. The first approach, which ties together existing DB and IR systems such as the one taken by SQL/MM [41], is not powerful enough since both systems are treated as black boxes. The second approach is based on extending DB systems with IR functionality, or vice versa. Jayavel argued that extending (R)DBMSs violates many assumptions hardwired into current database systems. For example, is author name a structured or text field? In addition, database operators have precise, well-defined semantics while in IR, even the query result is not well-defined. In addition, scoring is databases is an attribute tacked on as a relational column and it is not clear how it can generalize IR scoring. Jayavel also argued that extending an IR system would not work because IR systems provide little support for structured data. In addition, scoring does not take structure into account. Finally, Jayavel argued for a new system architecture that would eventually replace today's systems and that is based on three design principles: (i) *structural data independence* which should guarantee that users

[2] http://www.lexisnexis.com,
http://www.google.com,
http://thomas.loc.gov

can issue complex and keyword queries over structured and unstructured data, (ii) *generalized scoring* that operates over any mix of structured and unstructured data (e.g., XRank over HTML and XML [31]) and, (iii) *a flexible and powerful query language* that allows for arbitrary return results and scores (e.g., TeXQuery [3], XQuery Full-Text [59] and NEXI [34] languages).

2.3 Summary

1. *Potential data and applications* include LoC documents available at:
   http://www.loc.gov, a LoC search engine at http://thomas.loc.gov and customer support and Health care management.

2. *Research ideas:* (1) Realizing IR functionality in a DB system, and vice versa, provides a limited integration of their functionalities but could be a good solution for some applications where the main focus is on one kind of data or the other; (2) Standard end-user syntax (see XQuery Full-Text for XML search [59] but how about for non-XML data formats?); (3) Generalized scoring on structured and text content; (4) Approximate SQL, top-K ranking, parameterized ranking; (5) Approximate data integration and data cleaning; (6) New system architecture to unify DB and IR.

3. *Organizational ideas* include co-locating SIGIR and SIGMOD and participating to the INEX [34] and W3C FTTF efforts [59].

3. BIOS OF PANEL PARTICIPANTS

3.1 Moderator

Sihem Amer-Yahia is a researcher at AT&T Labs Research. She received her Ph.D. in Computer Science from the University Paris XI-Orsay and INRIA. She has been working on various issues related to XML query processing. Sihem is a co-editor of the XQuery Full-Text language specification [59] and use cases [58] published in September 2005 by the Full-Text Task Force in the W3C whose charter is to extend XQuery with full-text search and ranking capabilities. She is currently involved in the GalaTex project (www.galaxquery.org/galatex), a conformance implementation of XQuery Full-Text.

3.2 Panelists

bf Pat Case works for the Congressional Research Service at the U.S. Library of Congress. She is a Librarian who works as a search interface designer for the Legislative Information System – the Congress-access-only version of thomas.loc.gov/. Pat is a co-editor of the XQuery Full-Text language specification [59] and use cases [58] published in April 2005 by the Full-Text Task Force in the W3C whose charter is to extend XQuery with full-text search and ranking capabilities.

Thomas Rölleke attended from 1984-1986 a private computer school of former Nixdorf Computer. From 1986-1988, he was a management trainee and product consultant in the Unix marketing of Nixdorf Computer. In 1988, he started his studies in Computer Science, and obtained his MSc in 1994. In 1999, he obtained his PhD on "POOL: A probabilistic object-oriented logic for information retrieval". Since 2000, he has been working as strategic IT consultant for a leading online-bank, company directory, research fellow and lecturer at Queen Mary University in London (QMUL). Thomas Rölleke is currently the director of QMUL's first computer science spin-out. He holds a patent for a new SQL variant to support relevance-based retrieval in relational DBs. His research and activities are shaped by the vision that the integration of modern IR and DB technologies is an important step for increasing the productivity in building advanced information systems.

Jayavel Shanmugasundaram is an Assistant Professor at the Department of Computer Science at Cornell University. He obtained his Ph.D. degree from the University of Wisconsin, Madison. Prior to joining Cornell University, he spent two years at the IBM Almaden Research Center in San Jose, California. Jayavel's research interests include Internet data management, IR, and query processing in emerging system architectures. He is an invited expert to the W3C Full-Text Task Force, and is also the recipient of the NSF CAREER Award and an IBM Faculty Award.

Gerhard Weikum is a Research Director at the Max-Planck Institute of Computer Science in Saarbruecken, Germany. Earlier affiliations include the University of the Saarland in Germany, ETH Zurich in Switzerland, MCC in Austin, Texas, and, during a sabbatical, Microsoft Research in Redmond, Washington. Gerhard is co-author of more than 100 refereed publications, and he has written a textbook on Transactional Information Systems, published by Morgan Kaufmann. He received the 2002 VLDB ten-year award for his work on automatic tuning. His current research interests include intelligent search on semistructured data, combining DB technology with IR techniques, and "autonomic" peer-to-peer information management. Gerhard serves on the editorial boards of ACM TODS and IEEE CS TKDE, and he was the program committee chair for the 2004 SIGMOD conference in Paris.

4. REFERENCES

[1] S. Agrawal, S. Chaudhuri, G. Das, A. Gionis: Automated Ranking of Database Query Results. CIDR 2003.

[2] S. Al-Khalifa, C. Yu, H.V. Jagadish. Querying Structured Text in an XML Database. SIGMOD 2003.

[3] S. Amer-Yahia, C. Botev, J. Shanmugasundaram. TeXQuery: A Full-Text Search Extension to XQuery. WWW 2004.

[4] S. Amer-Yahia, N. Koudas, A. Marian, D. Srivastava, D. Toman Structure and Content Scoring for XML. To appear in VLDB 2005.

[5] S. Amer-Yahia, L. Lakshmanan, S. Pandit. FleXPath: Flexible Structure and Full-Text Querying for XML. SIGMOD 2004.

[6] R. Baeza-Yates, B. Ribeiro-Neto. Modern Information Retrieval. Addison-Wesley, 1999.

[7] R.A. Baeza-Yates, M.P. Consens. The Continued Saga of DB-IR Integration. Tutorial, VLDB 2004.

[8] A. Balmin, V. Hristidis, Y. Papakonstantinou. ObjectRank: Authority-Based Keyword Search in Databases. VLDB 2004.

[9] G. Bhalotia, A. Hulgeri, C. Nakhe, S. Chakrabarti, S. Sudarshan. Keyword Searching and Browsing in Databases using BANKS. ICDE 2002.

[10] H.M. Blanken, T. Grabs, H.-J. Schek, R. Schenkel, G. Weikum (Editors). Intelligent Search on XML Data, Applications, Languages, Models, Implementations, and Benchmarks. Springer, 2003.

[11] J. Bosak. The plays of Shakespeare in XML. http://www.oasis-open.org/cover/bosakShakespeare200.html

[12] J. M. Bremer, M. Gertz. XQuery/IR: Integrating XML Document and Data Retrieval. WebDB 2002.

[13] C. Botev, J. Shanmugasundaram. Context-Sensitive Keyword Search and Ranking for XML. WebDB 2005.

[14] E. W. Brown. Fast Evaluation of Structured Queries for Information Retrieval. SIGIR 1995.

[15] D. Carmel, Y. S. Maarek, M. Mandelbrod, Y. Mass, A. Soffer. Searching XML Documents via XML Fragments. SIGIR 2003.
[16] Soumen Chakrabarti. Breaking Through the Syntax Barrier: Searching with Entities and Relations. ECML 2004 and PKDD 2004.
[17] S. Chaudhuri, G. Das, V. Hristidis, G. Weikum. Probabilistic Ranking of Database Query Results. VLDB 2004.
[18] S. Chaudhuri, R. Ramakrishnan, G. Weikum. Integrating DB and IR Technologies: What is the Sound of One Hand Clapping? CIDR 2005.
[19] T. T. Chinenyanga, N. Kushmerick. Expressive and Efficient Ranked Querying of XML Data. WebDB 2001.
[20] E. F. Codd. A Relational Model of Data for Large Shared Data Banks. Commun. ACM 13(6): 377-387 (1970).
[21] E.F. Codd. Relational Completeness of Database Sublanguages. In R. Rustin (ed.), Database Systems, Prentice-Hall, 1972.
[22] S. Cohen, J. Mamou, Y. Kanza, Y. Sagiv. XSEarch: A Semantic Search Engine for XML. VLDB 2003.
[23] W.W. Cohen. Integration of Heterogeneous Databases Without Common Domains Using Queries Based on Textual Similarity. SIGMOD 1998.
[24] W. B. Croft. Language Models for Information Retrieval. Invited Talk. ICDE 2003.
[25] D. Florescu, D. Kossmann, I. Manolescu. Integrating Keyword Search into XML Query Processing. WWW 2000.
[26] DBLP in XML. http://dblp.uni-trier.de/xml/
[27] N. Fuhr, K. Grossjohann. XIRQL: An Extension of XQL for Information Retrieval. SIGIR 2001.
[28] N. Fuhr, K. Grossjohann. XIRQL: An XML query language based on information retrieval concepts. ACM TOIS 22(2), 2004.
[29] N. Fuhr, T. Rölleke. A Probabilistic Relational Algebra for the Integration of Information Retrieval and Database Systems. ACM TOIS 15(1), 1997.
[30] T. Grabs, K. Böhm, H.-J. Schek. PowerDB-IR - Information Retrieval on Top of a Database Cluster. CIKM 2001.
[31] L. Guo, F. Shao, C. Botev, J. Shanmugasundaram. XRANK: Ranked Keyword Search over XML Documents. SIGMOD 2003.
[32] Health Level Seven. http://www.hl7.org
[33] V. Hristidis, L. Gravano, Y. Papakonstantinou. Efficient IR-Style Keyword Search over Relational Databases. VLDB 2003.
[34] Initiative for the Evaluation of XML Retrieval. http://inex.is.informatik.uni-duisburg.de:2004/
[35] V. Kakade, P. Raghavan. Encoding XML in Vector Spaces. ECIR 2005.
[36] R. Kaushik, R. Krishnamurthy, J.F. Naughton, R. Ramakrishnan. On the Integration of Structure Indexes and Inverted Lists. SIGMOD 2004.
[37] B. Kimelfeld,Y. Sagiv. Efficient Engines for Keyword Proximity Search. WebDB 2005.
[38] M. Lalmas, T. Rölleke. Modelling Vague Content and Structure Querying in XML Retrieval with a Probabilistic Object-Relational Framework. FQAS 2004.
[39] Library of Congress. http://lcweb.loc.gov/crsinfo/xml/
[40] S. Liu, R. Shahinian, W. Chu. XML Vague Content and Structure (VCAS) Retrieval over Document-centric XML Collections. WebDB2005.
[41] J. Melton, A. Eisenberg. SQL Multimedia and Application Packages (SQL/MM). SIGMOD Record 30(4), 2001.
[42] A. Marian, S. Amer-Yahia, N. Koudas, D. Srivastava. Adaptive Processing of Top-k Queries in XML. ICDE 2005.
[43] J. Naughton, et al. The Niagara Internet Query System. IEEE Data Engineering Bulletin 24(2), 2001.
[44] N. Polyzotis, M. Garofalakis, Y. Ioannidis. Approximate XML Query Answers. SIGMOD 2004.
[45] S. Robertson. The probability ranking principle in IR. Journal of Documentation 33, 1977.
[46] A. Salminen. A Relational Model for Unstructured Documents. SIGIR 1987.
[47] G. Salton, M. J. McGill. Introduction to Modern Information Retrieval. McGraw-Hill, 1983.
[48] G. Salton, and A. Wong. A Vector Space Model for Automatic Indexing. Communications of the ACM 18, 1975.
[49] D. Shin, H. Jang, H. Jin. BUS: An Effective Indexing and Retrieval Scheme in Structured Documents. Proc. 3rd Int. Conf. on Dig. Lib., 1998.
[50] T. Schlieder. Schema-Driven Evaluation of Approximate Tree-Pattern Queries. EDBT 2002.
[51] A. Theobald, G. Weikum. Adding Relevance to XML. WebDB 2000.
[52] A. Theobald, G. Weikum. The Index-Based XXL Search Engine for Querying XML Data with Relevance Ranking. EDBT 2002.
[53] M. Theobald, R. Schenkel, G. Weikum. An Efficient and Versatile Query Engine for TopX Search. To appear in VLDB 2005.
[54] H. Turtle, B. Croft. Inference Networks for Document Retrieval. SIGIR 1990.
[55] F. Weigel, H. Meuss, K. U. Schulz, F. Bry. Content and Structure in Indexing and Ranking XML. WebDB 2004.
[56] The World Wide Web Consortium. XQuery 1.0: An XML Query Language. W3C Working Draft. http://www.w3.org/TR/xquery/.
[57] The World Wide Web Consortium. XML Path Language (XPath) 2.0. W3C Working Draft. http://www.w3.org/TR/xpath20/
[58] The World Wide Web Consortium. XQuery 1.0 and XPath 2.0 Full-Text Use Cases. W3C Working Draft. http://www.w3.org/TR/xmlquery-full-text-use-cases/
[59] The World Wide Web Consortium. XQuery 1.0 and XPath 2.0 Full-Text. W3C Working Draft. http://www.w3.org/TR/xquery-full-text/
[60] The World Wide Web Consortium. XQuery 1.0 and XPath 2.0 Functions and Operators. W3C Working Draft. http://www.w3.org/TR/xquery-operators/
[61] C. Zhang, J. Naughton, D. DeWitt, Q. Luo, G. Lohman. On Supporting Containment Queries in Relational Database Management Systems. SIGMOD 2001.
[62] E. Zimanyi. Query Evaluations in Probabilistic Relational Databases. Theoretical Computer Science, 1997.