

DB&IR: Both Sides Now (Extended Abstract)

Gerhard Weikum
Max-Planck Institute for Informatics
D-66123 Saarbruecken, Germany
weikum@mpi-inf.mpg.de

Categories and Subject Descriptors

H.2 [Database Management]: General; H.3 [Information Storage and Retrieval]: General

General Terms

Management

Keywords

DB&IR integration, database systems, information retrieval, XML, semantic search, Web knowledge

1. PAST

Database systems (DB) and information retrieval (IR) are two separate fields of computer science by historical accident. Both study concepts, models, and computational methods for managing large amounts of complex information, but thirty or forty years ago they started with very different application areas as major motivations and technology drivers: accounting systems (online reservations, banking, etc.) for DB, and library systems (bibliographic catalogs, patent collections, etc.) for IR. Thus, the two directions and their research communities emphasized very different aspects of information management: data consistency, precise query processing, and efficiency on the DB side [53], and text understanding, statistical ranking models, and user satisfaction on the IR side [35, 47].

Decades later, there is now rapidly growing awareness of the needs for integrating DB and IR technologies [3, 7, 14]. There have been various attempts of addressing this integration already ten years ago (e.g., [20, 29, 49]), but only recently important killer applications are emerging with really strong desire for an integrated DB&IR platform. From an IR viewpoint, digital libraries of all kinds are becoming very rich information repositories with documents augmented by metadata and annotations captured in semistructured data formats like XML [26]; enterprise search on intranet data can be seen as a specific variant of this theme. From a

DB viewpoint, application areas such as customer support, product and market research, or health-care management exhibit tremendous data growth, in terms of both structured and unstructured information, and at the same time become more and more mission-critical [11]. And completely new applications like Internet-based community management enjoy great popularity and pose interesting challenges [54].

Why has the envisioned DB&IR integration not yet happened? Why is it so difficult? At first glance, the key difference between DB and IR seems to lie in emphasizing different data types: numbers vs. text, or more precisely, structured records with numerical and categorical attributes in the DB world vs. unstructured or semistructured text documents in the IR world. But there are platforms that support both structured records and text (including the DBMS market leaders and a number of XML platforms), yet they do not provide truly satisfactory solutions with seamless integration. So the differences go much deeper. I believe that their root cause lies in the radically different notions of *users*:

- For DB systems and the DB research community, a user really is an application programmer who uses SQL, XQuery, or some APIs. In IR, on the other hand, a user is a non-technical human with cognitive capabilities and limitations. The consequences of these different user models are dramatic.
- DB systems expect the user to pose precise queries, and then aim to provide exact results in one shot and as fast as possible. IR systems understand queries as approximate, best-effort formulations of the user's information needs, and then aim to support an interactive process of data exploration, query rephrasing, and guidance towards the final results.
- Thus, DB systems view query processing as a *matching* task based on testing *logical* predicates, whereas IR views query processing as a *ranking* task based on *statistical* models.

In the last five years, the DB side has advanced on adding approximate and top-k query processing to its repertoire, and the IR side has paid more attention to semistructured data. Nevertheless, both communities still seem to underestimate the viewpoints and underestimate the benefits of the other side.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SIGMOD'07, June 12–14, 2007, Beijing, China.

Copyright 2007 ACM 978-1-59593-686-8/07/0006 ...\$5.00.

2. PRESENT

A contemporary research area that exemplifies the difficulties of DB&IR integration is XML IR. Non-schematic XML documents arise in a variety of situations:

- when digital libraries are loosely coupled into federated services,
- when combining data from many different schematic sources but without global schema or with only partial schema mappings, or
- when originally schematic documents are annotated and enriched by information-extraction methods or “social tagging” efforts.

Such XML data inevitably exhibits heterogeneous structures and tags and, therefore, cannot be adequately searched using matching-based DB query languages like XPath or XQuery. Often, queries either return too many or too few results. Rather the IR-style ranking paradigm is called for, with relaxable search conditions, various forms of similarity predicates on contents and structure, and quantitative relevance scoring. Ranked retrieval from multiple XML or other semistructured or even structured data sources may even be seen as a query-time approach to approximate information integration. Since the start of this millenium, significant research has gone into addressing these XML IR issues, and the early approaches like [16, 30, 59] have meanwhile converged to a consolidated state of the art (see, e.g., [8, 18, 21, 31, 36, 60] and references given there).

While most prior work on XML has focused on collections of trees, the option for XLinks and hyperlinks within and across semistructured documents motivates graph-oriented, extended approaches [19, 34]. This situation also arises with semistructured desktop data such as email, folder hierarchies, and other personal information [17, 23, 25, 48], and it is also related to keyword search on relational data graphs with database records as nodes and foreign-key relationships as edges [2, 10, 40]. Similarly, casting hyperlinked Web pages into XML so that complex queries can return groups of neighboring Web pages also leads to a graph IR problem. Finally, RDF triples naturally form complex graphs and thus call for graph querying as well [5, 32]. In all these settings, the result of a query is a subgraph spanned by nodes that approximately match and have high scores for the query’s elementary conditions. Finding the top-k, preferably compact, results may involve computationally hard problems related to Steiner trees, depending on how the query semantics and ranking models are defined. Notwithstanding recent progress (e.g., [22, 37, 42, 44]), graph IR continues to pose semantic as well as algorithmic challenges.

3. FUTURE

The Web has become one of mankind’s most impressive artifacts, without involving either one of the DB and IR research communities. We are currently witnessing various trends towards imposing more structure on both Web contents and search capabilities, bringing the Web closer to the DB world. Faceted search, vertical search, object search, and entity search are variations of the broader theme of finding, ranking, tracking, and analyzing semantic objects such as products (along with customer opinions), companies (and their market impacts), or researchers (and their

scholarly work) (see, e.g., [15, 38, 43, 50, 58] and references given there). Searching the Deep Web, the huge diversity of databases behind Web portals and query forms, requires mappings between (mostly schemaless) queries and record structures of potential target databases [13, 46]. In such Web-based but database-style settings, attempts for perfect entity recognition and perfect schema matching would be hopeless. Rather we must live with imperfect or noisy databases and entity-centric structures, which in turn mandates approximate search and ranking – a strong case for combined DB&IR methodology.

The theme of searching entities and relations rather than Web pages can be further expanded in scope and made even more ambitious by aiming to turn the entire Web into a gigantic knowledge base. A first, “smaller-scale” step could be to automatically turn Wikipedia into a database with explicit relations that contain all facts about people’s birthdates, professions, publications, awards, spouses and children, involvement in major events and their dates and places, and so on. Some ongoing projects (e.g., [24, 50]) pursue a similar direction for scholarly information, but the Wikipedia database would have a much larger scope and scale. With an explicit knowledge base of this kind, it should be possible to answer advanced knowledge queries such as: which physicist survived two world wars and died after all of his four children? This would go way beyond what Web search engines or even natural-language question-answering systems can provide today. Many questions of this kind could be answered more or less exactly, but many others will require reasoning about uncertainty, and thus need ranking in a DB&IR framework. On the full Web scale, information extraction technology could identify, connect, and organize even richer and many more pieces of knowledge, potentially leading to databases with facts about who invented or discovered what, which rivers run through which cities, or which enzymes trigger which biochemical processes, etc.

Of course, this dream of turning the Web into a database or knowledge base is not new at all. So why is *now* a good time to revive this vision and intensify the research towards making it reality? Various recent advances and strong trends enable this great opportunity that we are having now.

- First, information-extraction (IE) technology - entity recognition and learning relation patterns - has made enormous progress and become much more scalable in recent years [1, 41] and also much less dependent on human supervision [9, 27, 56]. Much of this progress comes from major advances in the underlying fields of natural language processing (NLP) and statistical learning, but there is also a much better understanding of algorithmic efficiency and how to engineer large-scale IE. To be clear, all these technologies will remain computationally expensive, but the gloomy picture of such issues being “AI-complete” and practically hopeless is gone.
- Second, there is a growing amount of “low-hanging fruit” that allows us to harvest knowledge without any rocket science. A large extent of this comes from the Web 2.0 trends, or more specifically, the human contributions to the emerging Social Web (aka. Human Semantic Web) in the form of tagging (and thus semantically annotating) Web pages, passages or phrases in pages, images, videos, etc. and creating so-called folk-

sonomies (e.g., [39]). Another big contributor is the strong proliferation of high-quality knowledge repositories with some explicit structure that is suitable for entity, relation, and topic recognition. Probably, Wikipedia is the best example. Although it is still primarily hyperlinked text, the link structure, the thematic categories to which articles are manually assigned, and the templates that are used for authoring certain types of articles (e.g., about music bands) provide enormous benefits for semantic tagging. Several recent projects have made excellent use of Wikipedia and similar sources for building explicit knowledge bases and connecting these with other sources (e.g., [6, 57]).

- Although the Semantic Web in its originally envisioned glorious form is still a very elusive goal, the vision itself has created a significant momentum towards creating ontologies and representing knowledge in more rigorous formats than text (see, e.g., [55, 61] and references given there). These include general-purpose ontologies and thesauri such as SUMO, OpenCyc, ConceptNet, or WordNet, as well as domain-specific ontologies and terminological taxonomies such as GeneOntology, SNOMED, or UMLS. While each of these collections alone may be viewed as fairly partial, connecting them and combining them with “softer” knowledge sources such as Wikipedia could be a powerful way of organizing more and more knowledge in rigorous representations that allow effective querying and reasoning. Richly annotated natural-language corpora such as multilingual thesauri, word-sense-tagged texts, or even representations in logic-based frames start becoming an interesting asset as well [28, 33, 45, 52].

While it is widely open how to best leverage these potential assets towards the envisioned automatic harvesting and organization of knowledge from the Web, both DB and IR technologies should play key roles. Combining the three major assets - large-scale information extraction, social tagging, and explicit knowledge sources like ontologies - requires statistical reasoning about uncertainty and well-founded ranking models in the IR tradition, but must equally pay great attention to efficiency and scalability of indexing and query processing, traditional DB virtues. An integrated DB&IR methodology and tool suite could play an even stronger role.

4. ACKNOWLEDGEMENTS

The above insights and opinions have benefited from many discussions with my collaborators in the area of DB&IR integration: Holger Bast, Srikanta Bedathur, Surajit Chaudhuri, Gautam Das, Norbert Fuhr, Vagelis Hristidis, Georgiana Ifrim, Gjergji Kasneci, Debapriyo Majumdar, Thomas Neumann, Raghu Ramakrishnan, Maya Ramanath, Ralf Schenkel, Fabian Suchanek, Anja Theobald, Martin Theobald, Christos Tryfonopoulos. However, all biases and flaws are solely mine.

5. REFERENCES

- [1] Eugene Agichtein, Sunita Sarawagi: Scalable Information Extraction and Integration. Tutorial Slides, KDD 2006. http://www.cs.columbia.edu/~eugene/kdd2006_tutorial/KDD06Tutorial.pdf
- [2] Sanjay Agrawal, Surajit Chaudhuri, Gautam Das: DBXplorer: A System for Keyword-Based Search over Relational Databases. ICDE 2002.
- [3] Sihem Amer-Yahia, Pat Case, Thomas Rölleke, Jayavel Shanmugasundaram, Gerhard Weikum: Report on the DB/IR panel at SIGMOD 2005. SIGMOD Record 34(4): 71-74, 2005.
- [4] Sihem Amer-Yahia, Jayavel Shanmugasundaram: XML Full-Text Search: Challenges and Opportunities. Tutorial Slides, VLDB 2005. <http://www.vldb2005.org/program/slides/fri/s1368-amer-yahia.ppt>
- [5] Kemafor Anyanwu, Angela Maduko, Amit Sheth, SPARQ2L: Towards Support For Subgraph Extraction Queries in RDF Databases. WWW 2007.
- [6] Sören Auer, Jens Lehmann: What have Innsbruck and Leipzig in common? Extracting Semantics from Wiki Content. ESWC 2007.
- [7] Ricardo Baeza-Yates, Mariano Consens: The Continued Saga of DB-IR Integration. Tutorial Slides, VLDB 2004. http://www.cs.toronto.edu/vldb04/protected/DB-IR_VLDB_1p.pdf
- [8] Ricardo Baeza-Yates, Mounia Lalmas: XML Information Retrieval. Tutorial Slides, SIGIR 2006. <http://www.dcs.qmul.ac.uk/~mounia/CV/XMLIR.pdf>
- [9] Michele Banko, Michael J. Cafarella, Stephen Soderland, Matt Broadhead, Oren Etzioni: Open Information Extraction from the Web. IJCAI 2007.
- [10] Gaurav Bhalotia, Arvind Hulgeri, Charuta Nakhe, Soumen Chakrabarti, S. Sudarshan: Keyword Searching and Browsing in Databases using BANKS. ICDE 2002.
- [11] Bishwaranjan Bhattacharjee, Joseph S. Glider, Richard A. Golding, Guy M. Lohman, Volker Markl, Hamid Pirahesh, Jun Rao, Robert Rees, Garret Swart: Impliance: A Next Generation Information Management Appliance. CIDR 2007.
- [12] Soumen Chakrabarti: Breaking Through the Syntax Barrier: Searching with Entities and Relations. ECML 2004.
- [13] Kevin Chen-Chuan Chang: Large-Scale Deep Web Integration: Exploring and Querying Structured Data on the Deep Web. Tutorial Slides, SIGMOD 2006. <http://www-sal.cs.uiuc.edu/~kcchang/talks/webitutorial-sigmod06-kcchang-jun06.ppt>
- [14] Surajit Chaudhuri, Raghu Ramakrishnan, Gerhard Weikum: Integrating DB and IR Technologies: What is the Sound of One Hand Clapping? CIDR 2005.
- [15] Tao Cheng, Kevin Chen-Chuan Chang: Entity Search Engine: Towards Agile Best-Effort Information Integration over the Web. CIDR 2007.
- [16] Taurai Tapiwa Chinenyanga, Nicholas Kushmerick: Expressive Retrieval from XML Documents. SIGIR 2001.
- [17] Paul-Alexandru Chirita, Stefania Costache, Wolfgang Nejdl, Raluca Paiu: Beagle++: Semantically Enhanced Searching and Ranking on the Desktop. ESWC 2006.
- [18] Jennifer Chu-Carroll, John M. Prager, Krzysztof Czuba, David A. Ferrucci, Pablo Ariel Duboué: Semantic Search via XML Fragments: a High-Precision Approach to IR. SIGIR 2006.

- [19] Sara Cohen, Yaron Kanza, Benny Kimelfeld, Yehoshua Sagiv: Interconnection Semantics for Keyword Search in XML. CIKM 2005.
- [20] William W. Cohen: Integration of Heterogeneous Databases Without Common Domains Using Queries Based on Textual Similarity. SIGMOD 1998.
- [21] Mariano P. Consens, Ricardo A. Baeza-Yates: Database and Information Retrieval Techniques for XML. ASIAN 2005.
- [22] Bolin Ding, Jeffrey Xu Yu, Shan Wang, Lu Qin, Xiao Zhang, Xuemin Lin: Finding Top-k Min-Cost Connected Trees in Databases. ICDE 2007.
- [23] Jens-Peter Dittrich, Marcos Antonio Vaz Salles: iDM: A Unified and Versatile Data Model for Personal Dataspace Management. VLDB 2006.
- [24] AnHai Doan, Raghu Ramakrishnan, Fei Chen, Pedro DeRose, Yoonkyong Lee, Robert McCann, Mayssam Sayyadian, Warren Shen: Community Information Management. IEEE Data Eng. Bull. 29(1): 64-72, 2006.
- [25] Xin Dong, Alon Y. Halevy: A Platform for Personal Information Management and Integration. CIDR 2005.
- [26] ERCIM News No. 66, Special Issue on European Digital Library, July 2006, http://www.ercim.org/publication/Ercim_News/enw66/
- [27] Oren Etzioni, Michael J. Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, Alexander Yates: Unsupervised Named-Entity Extraction from the Web: An Experimental Study. Artif. Intell. 165(1): 91-134, 2005.
- [28] Christiane Fellbaum (Ed.): WordNet: An Electronic Lexical Database. MIT Press, 1998.
- [29] Norbert Fuhr, Thomas Rölleke: A Probabilistic Relational Algebra for the Integration of Information Retrieval and Database Systems. ACM Trans. Inf. Syst. 15(1): 32-66, 1997.
- [30] Norbert Fuhr, Kai Großjohann: XIRQL: A Query Language for Information Retrieval in XML Documents. SIGIR 2001.
- [31] Norbert Fuhr, Mouna Lalmas (Eds.): Information Retrieval 8(4), Special Issue on INEX, December 2005.
- [32] Tim FURCHE, Benedikt Linse, François Bry, Dimitris Plexousakis, Georg Gottlob: RDF Querying: Language Constructs and Evaluation Methods Compared. Reasoning Web 2006.
- [33] Daniel Gildea, Daniel Jurafsky: Automatic Labeling of Semantic Roles. Computational Linguistics 28(3): 245-288, 2002.
- [34] Jens Graupmann, Ralf Schenkel, Gerhard Weikum: The SphereSearch Engine for Unified Ranked Retrieval of Heterogeneous XML and Web Documents. VLDB 2005.
- [35] David A. Grossman, Ophir Frieder: Information Retrieval: Algorithms and Heuristics. Springer, 2006.
- [36] Sudipto Guha, H. V. Jagadish, Nick Koudas, Divesh Srivastava, Ting Yu: Integrating XML Data Sources Using Approximate Joins. ACM Trans. Database Syst. 31(1): 161-207, 2006.
- [37] Hao He, Haixun Wang, Jun Yang, Philip Yu: BLINKS: Ranked Keyword Searches on Graphs. SIGMOD 2007.
- [38] Marti Hearst: Design Recommendations for Hierarchical Faceted Search Interfaces, SIGIR Workshop on Faceted Search, 2006.
- [39] Andreas Hotho, Robert Jäschke, Christoph Schmitz, Gerd Stumme: Information Retrieval in Folksonomies: Search and Ranking. ESWC 2006.
- [40] Vagelis Hristidis, Yannis Papakonstantinou: DISCOVER: Keyword Search in Relational Databases. VLDB 2002.
- [41] Panagiotis G. Ipeirotis, Eugene Agichtein, Pranay Jain, Luis Gravano: To Search or to Crawl?: Towards a Query Optimizer for Text-centric Tasks. SIGMOD 2006.
- [42] Varun Kacholia, Shashank Pandit, Soumen Chakrabarti, S. Sudarshan, Rushi Desai, Hrishikesh Karambelkar: Bidirectional Expansion For Keyword Search on Graph Databases. VLDB 2005.
- [43] Gjergji Kasneci, Fabian M. Suchanek, Georgiana Ifrim, Maya Ramanath, Gerhard Weikum: NAGA: Searching and Ranking Knowledge. Technical Report, Max-Planck Institute for Informatics, March 2007.
- [44] Benny Kimelfeld, Yehoshua Sagiv: Finding and Approximating Top-k Answers in Keyword Proximity Search. PODS 2006.
- [45] Hugo Liu, Push Singh: Commonsense Reasoning in and over Natural Language. KES 2004.
- [46] Jayant Madhavan, Shirley Cohen, Xin Luna Dong, Alon Y. Halevy, Shawn R. Jeffery, David Ko, Cong Yu: Web-Scale Data Integration: You can Afford to Pay as You Go. CIDR 2007.
- [47] Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze: Introduction to Information Retrieval. Cambridge University Press, 2007.
- [48] Einat Minkov, Andrew Ng, William W. Cohen: Contextual Search and Name Disambiguation in Email using Graphs. SIGIR 2006.
- [49] Gonzalo Navarro, Ricardo A. Baeza-Yates: Proximal Nodes: A Model to Query Document Databases by Content and Structure. ACM Trans. Inf. Syst. 15(4): 400-435, 1997.
- [50] Zaiqing Nie, Ji-Rong Wen, Wei-Ying Ma: Object-level Vertical Search. CIDR 2007.
- [51] Beng Chin Ooi, Bei Yu, Guoliang Li: One Table Stores All: Enabling Painless Free-and-Easy Data Publishing and Sharing. CIDR 2007.
- [52] Martha Palmer, Daniel Gildea, Paul Kingsbury: The Proposition Bank: An Annotated Corpus of Semantic Roles. Computational Linguistics, 31(1): 71-106, 2005.
- [53] Raghu Ramakrishnan, Johannes Gehrke: Database Management Systems. McGraw-Hill, 2002.
- [54] Raghu Ramakrishnan: Community Systems: The World Online. Keynote Slides, CIDR 2007, <http://www.cidrdb.org/cidr2007/slides/p39-ramakrishnan.ppt>
- [55] Steffen Staab, Rudi Studer: Handbook on Ontologies. Springer, 2004.
- [56] Fabian M. Suchanek, Georgiana Ifrim, Gerhard Weikum: Combining Linguistic and Statistical analysis to Extract Relations from Web Documents. KDD 2006.

- [57] Fabian Suchanek, Gjergji Kasneci, Gerhard Weikum: YAGO: A Core of Semantic Knowledge Unifying WordNet and Wikipedia. WWW 2007.
- [58] Dan Suci (Ed.): IEEE Data Engineering Bulletin 29(4), Special Issue on Web-Scale Data, Systems, and Semantics, December 2006.
- [59] Anja Theobald, Gerhard Weikum: Adding Relevance to XML. WebDB 2000.
- [60] Martin Theobald, Ralf Schenkel, Gerhard Weikum: An Efficient and Versatile Query Engine for TopX Search. VLDB 2005.
- [61] Gottfried Vossen, Miltiadis D. Lytras, Nick Koudas (Eds.): IEEE Trans. Knowl. Data Eng. 19(2), Special Issue on the Semantic Web Era, 2007.