The VLDB Journal (2008) 17:1–3
DOI 10.1007/s00778-007-0074-x

GUEST EDITORIAL

Introduction to the special issue on database and information retrieval integration

W. Bruce Croft · Hans-J. Schek

Published online: 6 November 2007

The goal of having a common platform for dealing with both structured and unstructured data is a longstanding one, going back to the 1960s. A number of approaches have been suggested, both from the database and information retrieval (IR) perspective, but the motivation for finding a solution or solutions that work has grown tremendously since the advent of very large-scale Web databases. Areas that were once the exclusive concerns of IR such as statistical inference and ranking, have now become important topics for database researchers and both communities have a common interest in providing efficient indexing and optimization techniques for Web-scale data. Exploiting document structure is a critical part of Web search and combining different sources of evidence effectively is an important part of many database applications. There are many possibilities for integration such as extending a database model to more effectively deal with probabilities, extending an IR model to handle more complex structures and multiple relations, or developing a unified model and system. Applications such as Web search, e-commerce, and data mining, provide the testbeds where these proposals can be evaluated and compared.

The papers in this special issue cover a range of topics related to database and IR integration. To provide some context, it is worth briefly reviewing some of the work that was done in the past, particularly in the more distant pre-Web days.

From an IR perspective, dealing with structure started in the 1970s with commercial search services such as MEDLINE and DIALOG that had Boolean field restrictions. In the 1970s and 80s, a number of papers described the implementation of IR systems using relational database systems (e.g. MacLeod 1979 [23]; Crawford 1981 [5]). Efficiency issues persisted with this approach until the 1990s (Defazio et al. 1995 [12]). Object management systems were also successfully used to support indexes in search engines (e.g. Brown et al. 1994 [3]). The 1990s also was the period when important work was done on developing probabilistic extensions of database models for IR applications. Fuhr and his colleagues described a probabilistic relational algebra (Fuhr 1990 [14]; Fuhr and Roelleke 1997 [16]) and a probabilistic datalog system (Fuhr 2000 [15]). Raghavan and his students also published a number of papers on the integration of retrieval models into a database (Deogun and Raghavan 1988 [13]; Saxton and Raghavan 1990 [25]). In the commercial world, text retrieval had become a standard function in database systems such as Oracle by the early 1990s, but the explosion of Web data and the growth of text-based Web applications later that decade made the ability to handle text effectively a critical part of most information systems.

Another important line of research in IR has been retrieval using structured documents. Early work in this area dealt with office documents (Croft et al. 1990 [8]) and document markup (Croft et al. 1992 [9]). Probabilistic models for IR that used structure-based evidence were also developed (Croft and Turtle 1992 [7]). More recently, there has been much research on retrieval with XML documents, motivated by the INEX initiative [1] that provides test databases, queries, and tasks.

Much of this research emphasizes the importance of a retrieval model. This is the formal description of the basis of combining evidence and ranking the database objects or documents for a particular query. Many, but not all, of these

W. Bruce Croft
University of Massachusetts Amherst, Amherst, USA
e-mail: croft@cs.umass.edu

H.-J. Schek (✉)
ETH Zurich, Zurich, Switzerland
e-mail: schek@inf.ethz.ch

[1] http://inex.is.informatik.uni-duisburg.de/

models are probabilistic, but all retrieval models are evaluated using rigorous experiments with real test collections. As more papers on databases and IR are published in the database literature, it is important that the retrieval models used are state-of-the-art, and that they are evaluated appropriately. For example, many papers are published in the database area using "tf.idf" ranking as the canonical IR method. IR practitioners would consider the ranking techniques used in these papers to be more heuristic and less effective than the current techniques. In terms of query and indexing optimization, there is also considerable literature in the IR area, dating back to the 1980s. Given that the designers of search engines have had to deal with the problem of ranking results from large databases for many years, this should not be surprising, but this work is often overlooked by the database community. Instead of reviewing these papers here, the recent survey by Zobel and Moffat (2006) [32] provides an excellent overview.

From the database perspective, the early 1980s saw a number of papers that discussed whether and how IR could be supported by DBMSs. Parallel to IR efforts that tried to build on top of the emerging relational databases as described above, researchers coming more from a database background questioned the usefulness of relational databases for IR and made proposals for extensions. The main concerns were the overall architecture, the data model and its query language, the access and indexing mechanisms, and transaction management.

On the architecture side, IR preprocessor versus hybrid ("middleware") versus unified approaches for a DB&IR integration were discussed (e.g. Schek 1980 [26]; Biller 1982 [2]; Dadam et al. 1986 [11]). Considering the data model, there was broad agreement already in the early 1980s (long before XML became a success) that the rigid first normal form condition and the related query language were too narrow. On the more conservative side, an extension by suitable procedures encapsulated in ADTs was proposed (e.g. Stonebraker et al. 1987 [29]). A slightly more radical approach by allowing nested relations was proposed by others (e.g. the NF2 model by Schek and Pistor 1982 [28]). In order to cope with positional information in documents, Güting and Zicari (1989) [20] and Güting et al. (1989) [21] proposed a model that supports nested sequences. Even more radical was the model in the AIM-P prototype that supported nested sets, records and sequences in all combinations (e.g. Dadam et al. 1986 [11]; Pistor and Dadam 1989 [24]).

On the access and transaction side, advantages were seen when using stable fragments of text (variable lengths n-grams and frequent term sequences) for indexing (e.g. Schek and Pistor 1982 [28]). Such an index can be adapted to the usage and, due to the stability of the index keys, the update effort can be kept smaller. Nevertheless, inserting or updating documents in parallel with the readers was considered difficult due to the database mechanisms for concurrency control and recovery. Early proposals for a solution include a deferred update concept together with differential files (Schek 1980 [26]). A generalization of the DAG locking of the System R prototype was proposed in Dadam et al. (1983) [10]. More radically, multi-level transactions were proposed afterward with the idea of considering the relational database as a storage engine and splitting an IR transaction into several shorter SQL transactions that can be committed independently without sacrificing correctness (Schek 1984 [27]; Weikum and Schek 1984 [31]).

Due to the maturity of relational database system implementations and their availability on multiprocessor systems, direct support of textual fields in relational databases was re-visited and evaluated again in the 1990s. Of particular interest is the method described first in Grossman (1992) [18]. He proposed to place the terms of an IR vector space query into a (small) query relation. This leads to a single join only between this query relation and the index term relation, regardless of the number of terms in a query, and therefore to a surprisingly good performance. This method and its evaluation of a parallel relational database system is described in Grossmann et al. (1997) [19]. Comparisons between a native IR system and IR implemented on top of relational databases and database clusters are found in, for example, Kaufmann and Schek (1995) [22] and Grabs et al. (2004) [17].

Recently, the awareness of the need for a close integration of IR and DB functionality has grown considerably as expressed by keynote speeches and panels at the mainstream database conferences (e.g. Weikum 2007 [30]; Amer-Yahia 2005 [1]; Croft 2006 [6]). The paper by Chaudhuri et al. (2005) [4] that describes some approaches to DB/IR integration has also received considerable attention.

Given the context of this earlier work, it is much easier to see how the papers in this special issue are related. The paper by Roelleke et al. describes new developments in a probabilistic relational algebra for retrieval models that follow his earlier work with Fuhr. Schmitt's paper describes how a new modeling approach based on quantum logic that is being used in IR may be used as the basis for an integrated data base and IR system. Lau et al.'s paper describes an approach to search XML data, and Schenkel et al.'s paper describes optimization techniques for an XML retrieval system. Koutrika et al. show how keyword queries can be used on top of relational databases that result in relevant, related subsets of the relations. The performance aspect of integration is the focus of Heman et al.'s paper. They propose a sparse array database on top of the relational Monet DB and show how IR queries can be supported efficiently.

The papers in this issue have been reviewed and selected by prominent experts in this field. The authors were very responsive to the feedback from the reviewers and made numerous revisions. We would like to thank the authors and the reviewers for their efforts.

References

1. Amer-Yahia, S., Case, P., Roelleke, T., Shanmugasundaram, J., Weikum, G.: Report on the DB/IR panel at SIGMOD 2005. SIGMOD Record **34**(4), 71–74 (2005)
2. Biller, H.: On the architecture of a system integrating data base management and information retrieval. Proc. ACM SIGIR 80–97 (1982)
3. Brown, E.W., Callan, J.P., Croft, W.B., Moss, J.E.B.: Supporting full-text information retrieval with a persistent object store. In: Proceedings of the 4th International Conference on Extending Database Technology (EDBT), pp. 363–378 (1994)
4. Chaudhuri, S., Ramakrishnan, R., Weikum, G.: Integrating DB and IR Technologies: What is the Sound of One Hand Clapping? In: Proceedings of the 2nd Biennial Conference on Innovative Data Systems Research (CIDR 05), pp. 1–12 (2005)
5. Crawford, R.: The relational model in information retrieval. J. Am. Soc. Inf. Sci. **32**, 51–64 (1981)
6. Croft, W.B.: Why can't we all get along? Keynote at 3rd Database/Information Retrieval Day, New York University (2006)
7. Croft, W.B., Turtle, H.: Retrieval of complex objects. Proc. EDBT **92**, 217–229 (1992)
8. Croft, W.B., Krovetz, R., Turtle, H.: Interactive retrieval of complex documents. Inf. Process. Manage. **26**(5), 593–613 (1990)
9. Croft, W.B., Smith, L., Turtle, H.: A loosely coupled integration of a text retrieval system and an object-oriented database system. Proc. ACM SIGIR **92**, 223–232 (1992)
10. Dadam, P., Pistor, P., Schek, H.-J.: A Predicate oriented locking approach for integrated information systems. Proc. IFIP Congr. 763–768 (1983)
11. Dadam, P., Küspert, K., Andersen, F., Blanken, H.M., Erbe, R., Günauer, J., Lum, V.Y., Pistor, P., Walch, G.: A DBMS prototype to support extended NF2 relations: an integrated view on flat tables and hierarchies. Proc. SIGMOD Conf. 356–367 (1986)
12. Defazio, S., Daoud, A., Smith, L., Srinivasan, J., Croft, W.B., Callan, J.: Integrating IR and RDBMS Using Cooperative Indexing, ACM Eighteenth International Conference on Research and Development in Information Retrieval (SIGIR), pp. 84–92 (1995)
13. Deogun, J.S., Raghavan, V.V.: Integration of information retrieval and database management systems. Inf. Process. Manage. **24**(3), 303–313 (1988)
14. Fuhr, N.: A Probabilistic Framework for Vague Queries and Imprecise Information in Databases. VLDB Proc. 696–707 (1990)
15. Fuhr, N.: Probabilistic datalog: implementing logical information retrieval for advanced applications. J. Am. Soc. Inf. Sci. **51**(2), 95–110 (2000)
16. Fuhr, N., Rölleke, T.: A probabilistic relational algebra for the integration of information retrieval and database systems. ACM Trans. Info. Syst. **15**, 32–66 (1997)
17. Grabs, T., Böhm, K., Schek, H.-J.: PowerDB-IR—scalable information retrieval and storage with a cluster of databases. Knowl. Inf. Syst. **6**(4), 465–505 (2004)
18. Grossman, D.A.: Using the relational model and part-of-speech tagging to implement text relevance. In: Proceedings of the First International Conference on Information and Knowledge Management (CIKM '92) (1992)
19. Grossman, D.A., Frieder, O., Holmes, D.O., Roberts, D.C.: Integrating structured data and text: a relational approach. J. Am. Soc. Inf. Sci. **48**(2) (1997)
20. Güting, R.H., Zicari, R. : An Introduction to the nested sequence of tuples data model and algebra. In: Abiteboul, S., Fischer, P.C., Schek, H.-J. (eds.) Nested Relations and Complex Objects in Databases. LNCS, vol. 361, Springer, Heidelberg (1989)
21. Güting, R.H., Zicari, R., Choy, D.M.: An algebra for structured office documents. ACM Trans. Off. Inf. Syst. **7**, 123–157 (1989)
22. Kaufmann, H., Schek, H.-J.: Text search using database systems revisited—some experiments. Proc. BNCOD 204–225 (1995)
23. MacLeod, I.A.: SEQUEL as a language for document retrieval. J. Am. Soc. Inf. Sci. **30**(2), 243–249 (1979)
24. Pistor, P., Dadam, P. : The advanced information management prototype. In: Abiteboul, S., Fischer, P.C., Schek, H.-J. (eds.) Nested Relations and Complex Objects in Databases. LNCS, vol. 361, Springer, Heidelberg (1989)
25. Saxton, L.V., Raghavan, V.V.: Design of an integrated information retrieval/database management system. IEEE Trans. Knowl. Data Eng. **2**(2), 210–219 (1990)
26. Schek H.-J.: Methods for the administration of textual data in database systems. Proc. ACM SIGIR 218–235 (1980)
27. Schek, H.-J.: Nested transactions in a combined IRS-DBMS architecture. Proc. ACM SIGIR 55–70 (1984)
28. Schek, H.-J., Pistor, P.: Data structures for an integrated data base management and information retrieval system. Proc. VLDB 197–207 (1982)
29. Stonebraker, M., Anton, J., Hanson, E.N.: Extending a database system with procedures. ACM Trans. Database Syst. **12**(3), 350–376 (1987)
30. Weikum, G., DB& IR: Both Sides Now, Keynote at SIGMOD'07. Beijing, June 14 (2007)
31. Weikum, G., Schek, H.-J.: Architectural issues of transaction management in multi-layered systems. Proc. VLDB 454–465 (1984)
32. Zobel, J., Moffat, A.: Inverted files for text search engines. ACM Comput. Surv. **38**(2), 1–56 (2006)