Database and Information Retrieval Techniques for XML

Mariano P. Consens[1] and Ricardo Baeza-Yates[2]

[1] University of Toronto
Toronto, Canada
consens@cs.toronto.edu
[2] ICREA – Univ. Pompeu Fabra
Barcelona, Spain
ricardo.baeza@upf.edu

1 Overview

The world of data has been developed from two main points of view: the structured relational data model and the unstructured text model. The two distinct cultures of databases and information retrieval now have a natural meeting place in the Web with its semi-structured XML model. As web-style searching becomes an ubiquitous tool, the need for integrating these two viewpoints becomes even more important.

This tutorial[3] will provide an overview of the different issues and approaches put forward by the Information Retrieval and the Database communities and survey the DB-IR integration efforts with a focus on techniques applicable to XML retrieval. A variety of application scenarios for DB-IR integration will be covered, including examples of current industrial tools.

2 Tutorial Content

The tutorial consists of two parts: the first part will cover the problem space (basic concepts, requirements, models) and the second part the solution space (approaches and techniques).

The major topics covered together with specific bibliographic references are listed below.

Introduction. Types of data, DB & IR views, Applications, Tokenization, Web Challenges ([1–3]).

Requirements for DB-IR. Motivation, Data and Query Requirements, Sample Use Cases ([4, 5]).

[3] Earlier versions of this tutorial have been given at VLDB 2004 and SIGIR 2005.

Semi-structured text models. XPat and XQuery, Full-text extensions to XQuery, Structured text models, Query algebras ([6–12]).

DB Approaches. IR on Relational Data and IR on XML: keyword search, full query language with extensions, algebras and evaluation ([13–18, 17, 19–27]).

IR and Hybrid Approaches. Retrieval models, Ranking, Evaluation ([28–38]).

Query Processing. XML Processing Algorithms (summaries, indexes), Query Optimization ([39–73]).

Open Problems. A discussion of research problems in the area.

Additional Reading. The following proceedings are relevant to the tutorial material.

- Proceedings of the ACM SIGIR Workshops on XML and Information Retrieval (edited by Yoelle Maarek *et al.*), 2002 & 2002.
- Proceedings of the workshops of the Initiative for the Evaluation of XML Retrieval (INEX) (edited by N. Fuhr, G. Kazai, M. Lalmas *et al*), 2002-2004.
- Special JASIST issue on XML and IR, 53(6): 2002. Edited by Ricardo Baeza-Yates, David Carmel, Yoelle Maarek, and Aya Sofer.
- Proceedings of First International Workshop on XQuery Implementation, Experience and Perspectives (XIME-P 2004) (edited by Ioana Manolescu and Yannis Papakonstantinou), June 2004.
- Proceedings of First and Second International XML Database Symposium (XSym), 2003-2004.
- Proceedings of Joint Workshop on XML and DB-IR Integration (edited by Ricardo Baeza-Yates, Yoelle Maarek, Thomas Roelleke, and Arjen P. de Vries), SIGIR 2004, Sheffield, 2004.

References

1. Salton, G.: Automatic information organization and retrieval. McGraw-Hill, New York, USA (1968)
2. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. Addison-Wesley, Harlow, UK (1999)
3. Crestani, F., Lalmas, M., van Rijsbergen, C.J., Campbell, I.: Is this document relevant? ... probably: A survey of probabilistic models in information retrieval. ACM Computing Surveys **30** (1998) 528–552
4. W3C: XQuery and XPath full-text requirements (2003) W3C Working Draft, http://www.w3.org/TR/xmlquery-full-text-requirements.

5. W3C: XQuery and XPath full-text use cases (2003) W3C Working Draft, http://www.w3.org/TR/xmlquery-full-text-use-cases.
6. Salminen, A., Tompa, F.W.: PAT expressions: An algebra for text search. Acta Linguistica Hungarica **41** (1993) 277–306
7. Consens, M., Milo, T.: Algebras for querying text regions. In: Proceedings of the Symposium on Principles of Database Systems, San Jose, California, USA (1995) 11–22
8. Clarke, C., Cormack, G., Burkowski, F.: An algebra for structured text search and a framework for its implementation. The Computer Journal **38** (1995) 43–56
9. Navarro, G., Baeza-Yates, R.: Integrating content and structure in text retrieval. SIGMOD Record **25** (1996) 67–79
10. Navarro, G., Baeza-Yates, R.: Proximal nodes: A model to query document databases by contents and structure. ACM Transactions on Information Systems **15** (1997) 401–435
11. Lee, Y.K., Yoo, S.J., Yoon, K., Berra, P.B.: Index structures for structured documents. In: Proceedings of the 1st ACM International Conference on Digital Libraries. (1996) 91–99
12. Navarro, G., Baeza-Yates, R.A.: Proximal nodes: A model to query document databases by content and structure. ACM Transactions on Information Systems (TOIS) **15** (1997) 400–435
13. Goldman, R., Shivakumar, N., Venkatasubramanian, S., Garcia-Molina, H.: Proximity search in databases. In: Proceedings of the 24th International Conference on Very Large Data Bases. (1998) 26–37
14. Florescu, D., Kossmann, D., Manolescu, I.: Integrating keyword search into XML query processing. In: Proceedings of International World Wide Web Conference. (2000)
15. Kanza, Y., Sagiv, Y.: Flexible queries over semistructured data. In: Proceedings of the Symposium on Principles of Database Systems. (2001) 40–51
16. Agrawal, S., Chaudhuri, S., Das, G.: DBXplorer: A system for keyword-based search over relational databases. In: Proceedings of International Conference on Data Engineering. (2002)
17. Bhalotia, G., Hulgeri, A., Nakhey, C., Chakrabarti, S., Sudarshan, S.: Keyword searching and browsing in databases using BANKS. In: Proceedings of International Conference on Data Engineering. (2002)
18. Hristidis, V., Papakonstantinou, Y.: DISCOVER: Keyword search in relational databases. In: Proceedings of the International Conference on Very Large Data Bases. (2002)
19. Amer-Yahia, S., Cho, S., Srivastava, D.: Tree pattern relaxation. In: Proceedings of Conference on Extending Database Technology. (2002) 496–513
20. Amer-Yahia, S., Fernandez, M., Srivastava, D., Xu, Y.: Pix: exact and approximate phrase matching in xml. In: Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data. (2003) 664–664
21. Kabra, N., Ramakrishnan, R., Ercegovac, V.: The QUIQ engine: A hybrid IR-DB system. In: Proceedings of the 19th International Conference on Data Engineering. (2003) 741
22. Amer-Yahia, S., Koudas, N., Srivastava, D.: Approximate matching in xml. In: Proceedings of the 19th International Conference on Data Engineering. (2003) 803
23. Hristidis, V., Papakonstantinou, Y., Balmin, A.: Keyword proximity search on XML graphs. In: Proceedings of International Conference on Data Engineering. (2003)

24. Guo, L., Shao, F., Botev, C., Shanmugasundaram, J.: XRANK: Ranked keyword search over XML documents. In: Proceedings of ACM SIGMOD International Conference on Management of Data. (2003)
25. Cohen, S., Mamou, J., Kanza, Y., Sagiv, Y.: XSearch: a semantic search engine for XML. In: Proceedings of the 29th International Conference on Very Large Data Bases. (2003)
26. Hristidis, V., Gravano, L., Papakonstantinou, Y.: Efficient IR-style keyword search over relational databases. In: Proceedings of the International Conference on Very Large Data Bases. (2003)
27. Amer-Yahia, S., Lakshmanan, L.V.S., Pandit, S.: FleXPath: Flexible structure and full-text querying for XML. In: Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data. (2004) 83–94
28. Luk, R.: A survey of search engines for XML documents. In: SIGIR Workshop on XML and IR. (2000)
29. Fuhr, N., Grobjohann, K.: XIRQL: An extension of XQL for information retrieval. In: ACM SIGIR Workshop on XML and Information Retrieval. (2000) 11–17
30. Theobald, A., Weikum, G.: Adding relevance to XML. In: Proceedings of International Workshop on the Web and Databases. (2000) 35–40
31. Fuhr, N., Grobjohann, K.: XIRQL: A query language for information retrieval in XML documents. In: Proceedings of ACM SIGIR International Conference on Research and Development in Information Retrieval. (2001) 172–180
32. Chinenyanga, T.T., Kushmerick, N.: Expressive and efficient ranked querying of XML data. In: Proceedings of International Workshop on the Web and Databases. (2001)
33. Theobald, A., Weikum, G.: The index-based XXL search engine for querying XML data with relevance ranking. In: Proceedings of Conference on Extending Database Technology. (2002) 477–495
34. Chinenyanga, T.T., Kushmerick, N.: An expressive and efficient language for XML information retrieval. Journal of the American Society for Information Science and Technology **53** (2002) 438–453
35. Grabs, T., Schek, H.J.: Flexible information retrieval from XML with PowerDB-XML. In: Proceedings of the Third INEX Workshop. (2003)
36. Mass, Y., Mandelbrod, M., Amitay, E., Carmel, D., Maarek, Y., Soffer, A.: JuruXML - an XML retrieval system at INEX 02. In: Proceedings of the First INEX Workshop. (2002)
37. Fuhr, N., Grobjohann, K.: XIRQL: An XML query language based on information retrieval concepts. ACM Trans. Inf. Syst. **22** (2004) 313–356
38. Schenkel, R., Theobald, A., Weikum, G.: Semantic similarity search on semistructured data with the XXL search engine. Information Retrieval **8** (2005) 521–545
39. Goldman, R., Widom, J.: Dataguides: Enabling query formulation and optimization in semistructured databases. In: Proceedings of the 23rd International Conference on Very Large Data Bases. (1997) 436–445
40. Nestorov, S., Ullman, J.D., Wiener, J.L., Chawathe, S.S.: Representative objects: Concise representations of semistructured, hierarchial data. In: Proceedings of the 13th International Conference on Data Engineering. (1997) 79–90
41. Milo, T., Suciu, D.: Index structures for path expressions. In: Proceedings of the 7th International Conference on Database Theory. (1999) 277–295
42. Cooper, B., Sample, N., Franklin, M.J., Hjaltason, G.R., Shadmon, M.: A fast index for semistructured data. In: Proceedings of the 27th International Conference on Very Large Data Bases. (2001) 341–350

43. Natsev, A., Chang, Y.C., Smith, J.R., Li, C.S., Vitter, J.S.: Supporting incremental join queries on ranked ranked inputs. In: Proceedings of the International Conference on Very Large Data Bases. (2001)
44. Fagin, R., Lotem, A., Naor, M.: Optimal aggregation algorithms for middleware. In: Proceedings of the Symposium on Principles of Database Systems. (2001)
45. Rizzolo, F., Mendelzon, A.O.: Indexing XML data with ToXin. In: Proceedings of 4th International Workshop on the Web and Databases. (2001) 49–54
46. Li, Q., Moon, B.: Indexing and querying XML data for regular path expressions. In: Proceedings of the 27th International Conference on Very Large Data Bases. (2001) 361–370
47. Kaushik, R., Bohannon, P., Naughton, J.F., Korth, H.F.: Covering indexes for branching path queries. In: Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data. (2002) 133–144
48. Chung, C.W., Min, J.K., Shim, K.: APEX: An adaptive path index for XML data. In: Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data. (2002) 121–132
49. Kaushik, R., Shenoy, P., Bohannon, P., Gudes, E.: Exploiting local similarity for indexing paths in graph-structured data. In: Proceedings of the 18th International Conference on Data Engineering. (2002) 129–140
50. Kaushik, R., Bohannon, P., Naughton, J.F., Shenoy, P.: Updates for structure indexes. In: Proceedings of the 28th International Conference on Very Large Data Bases. (2002) 239–250
51. Al-Khalifa, S., Jagadish, H.V., Patel, J.M., Wu, Y., Koudas, N., Srivastava, D.: Structural joins: A primitive for efficient XML query pattern matching. In: Proceedings of the 18th International Conference on Data Engineering. (2002) 141–
52. Chien, S.Y., Vagena, Z., Zhang, D., Tsotras, V.J., Zaniolo, C.: Efficient structural joins on indexed XML documents. In: Proceedings of the 28th International Conference on Very Large Data Bases. (2002) 263–274
53. Bruno, N., Koudas, N., Srivastava, D.: Holistic twig joins: Optimal XML pattern matching. In: Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data. (2002) 310–321
54. Hristidis, V., Papakonstantinou, Y.: Algorithms and applications for answering ranked queeries using ranked views. In: Proceedings of the International Conference on Very Large Data Bases. (2003)
55. Ilyas, I.F., Aref, W.G., Elmagarmid, A.K.: Supporting top-k join queries in relational databases. In: Proceedings of the International Conference on Very Large Data Bases. (2003)
56. Bremer, J.M.: Next-Generation Information Retrieval: Integrating Document and Data Retrieval Based on XML. PhD thesis, Department of Computer Science, University of California at Davis (2003)
57. Bremer, J.M., Gertz, M.: An efficient XML node identification and indexing scheme. Technical Report CSE-2003-04, Department of Computer Science, University of California at Davis (2003)
58. Chen, Z., Jagadish, H.V., Lakshmanan, L.V.S., Paparizos, S.: From tree patterns to generalized tree patterns: On efficient evaluation of XQuery. In: Proceedings of the 29th International Conference on Very Large Data Bases. (2003) 237–248
59. Al-Khalifa, S., Yu, C., Jagadish, H.V.: Querying structured text in an XML database. In: Proceedings of ACM SIGMOD International Conference on Management of Data. (2003)

60. Qun, C., Lim, A., Ong, K.W.: D(K)-index: An adaptive structural summary for graph-structured data. In: Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data. (2003) 134–144
61. Ramanan, P.: Covering indexes for XML queries: Bisimulation - simulation = negation. In: Proceedings of the 29th International Conference on Very Large Data Bases. (2003) 165–176
62. Zezula, P., Amato, G., Debole, F., Rabitti, F.: Tree signatures for XML querying and navigation. In: First International XML Database Symposium, XSym 2003. (2003) 149–163
63. Wang, H., Park, S., Fan, W., Yu, P.S.: ViST: A dynamic index method for querying XML data by tree structures. In: Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data. (2003) 110–121
64. Jiang, H., Wang, W., Lu, H., Yu, J.X.: Holistic twig joins on indexed XML documents. In: Proceedings of the 29th International Conference on Very Large Data Bases. (2003) 273–284
65. Jiang, H., Lu, H., Wang, W., Ooi, B.C.: XR-Tree: Indexing XML data for efficient structural joins. In: Proceedings of the 19th International Conference on Data Engineering. (2003) 253–263
66. Li, Q., Moon, B.: Partition based path join algorithms for XML data. In: Proceedings of the 14th International Conference on Database and Expert Systems Applications, DEXA 2003. (2003) 160–170
67. Weigel, F., Meuss, H., Bry, F., Schulz, K.U.: Content-aware dataguides: Interleaving IR and DB indexing techniques for efficient retrieval of textual XML data. In: Proceedings of the 26th European Conference on IR Research, ECIR 2004. (2004) 378–393
68. Kaushik, R., Krishnamurthy, R., Naughton, J.F., Ramakrishnan, R.: On the integration of structure indexes and inverted lists. In: Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data. (2004) 779–790
69. Amato, G., Debole, F., Rabitti, F., Savino, P., Zezula, P.: A signature-based approach for efficient relationship search on XML data collections. In: Second International XML Database Symposium, XSym 2004. (2004) 82–96
70. Rao, P., Moon, B.: PRIX: Indexing and querying XML using Prüfer sequences. In: Proceedings of the 20th International Conference on Data Engineering. (2004) 288–300
71. Vagena, Z., Moro, M.M., Tsotras, V.J.: Efficient processing of XML containment queries using partition-based schemes. In: Proceedings of the 8th International Database Engineering and Applications Symposium, IDEAS 2004. (2004) 161–170
72. Wang, H., Meng, X.: On the sequencing of tree structures for XML indexing. In: Proceedings of the 21st International Conference on Data Engineering. (2005)
73. Bremer, J.M., Gertz, M.: Next-generation information retrieval. VLDB Journal (2006) To appear.