Keyword-based Search and Exploration on Databases

Yi Chen [#1], Wei Wang [*2], Ziyang Liu [#3]

[#] *Arizona State University, USA*
[1] yi@asu.edu
[3] ziyang.liu@asu.edu

[*] *University of New South Wales, Australia*
[2] weiw@cse.unsw.edu.au

***Abstract*—Empowering users to access databases using simple keywords can relieve users from the steep learning curve of mastering a structured query language and understanding complex and possibly fast-evolving data schemas. In this tutorial, we give an overview of the state-of-the-art techniques for supporting keyword-based search and exploration on databases. Several topics will be discussed, including query result definition, ranking functions, result generation and top-k query processing, snippet generation, result clustering, result comparison, query cleaning and suggestion, performance optimization, and search quality evaluation. Various data models will be discussed, including relational data, XML data, graph-structured data, data streams, and workflows. Finally we identify the challenges and opportunities for future research to advance the field.**

I. Introduction

Web search engines are very successful for searching textual documents, images, and videos. On the other hand, there are also vast collections of structured and semi-structured data both on the Web and in enterprises, referred to as "databases" in this tutorial, such as relational databases, XML, data extracted from text documents, workflows, etc. Traditionally, to access these resources, users have to learn structured query languages, such as SQL or XQuery. Besides, users also need to know the schemas of the data, which are most likely complex, fast-evolving, or even unavailable in Web applications. Given the prevalence of web search engines, a natural question to ask is whether we can empower users to effectively search and explore databases using keyword queries.

There are several immediate advantages of this keyword query-based approach to search and explore the databases. First, it can relieve casual users from the steep learning curve of studying structured query languages and data schemas when accessing structured data. Second, it allows users to easily access heterogeneous databases. For instance, for websites with database back-ends, this approach provides a more flexible search method than the existing solution that relied on a fixed set of pre-built template queries. Third, ideally the result of a keyword search over databases will automatically assemble relevant pieces of data that are in different locations but are inter-connected and collectively relevant to the query. Thus unlike the results to a structured query, the results to a keyword query may reveal interesting or unexpected relationships hidden in the databases. Furthermore, the rich meta-information in databases holds enormous promise for achieving better search quality and enabling more effective data analysis compared with searching unstructured textual documents. Making database searchable will increase the database usability, and substantially increase the information volume that a user can access, thus making significant impact to people's lives.

Due to substantial benefits of supporting keyword search on structured data, it becomes a mainstream in database research and development. Researchers from different disciplines (e.g., information retrieval and theoretical computer science) are joining the workforce to tackle various challenges in supporting keyword search on structured data. Major database research laboratories, such as Microsoft and IBM, are working in this area [1], [71], [79]. The workshop series *Keyword Search on Structured Data (KEYS)* [29] were held in conjunction with ACM SIGMOD/PODS.

The mission of supporting keyword search on structured data is well aligned with recent keynotes in major database conferences [21], [27], [68], [78]. This tutorial, and an earlier tutorial offered in SIGMOD 2009 [11] [1], as well as related tutorials "XML Full-Text Search: Challenges and Opportunities" [2] by Amer-Yahia and Shanmugasundaram in VLDB 2005, and "Keyword querying and Ranking in Databases"[8] by Chaudhuri and Das in VLDB'09, provide overview of research advances that integrate database and information retrieval technologies.

II. Tutorial Outline

The objective of this tutorial is to provide a systematic and well-organized overview of the state-of-the-art in supporting keyword-based search and exploration on databases, outline the problem space in this area, introduce representative techniques that address different aspects of the problem, and discuss further challenges and promising directions for future work.

[1] available at http://www.public.asu.edu/~ychen127/keyword_sigmod09_tutorial.pptx

We will give an overview of the core problems on processing keyword queries on databases, including query result definition and result generation [5], [6], [13], [19], [20], [22], [24], [28], [30], [31], [32], [42], [37], [47], [45], [52], [55], [63], [70], [67], [81], [82], ranking functions [5], [13], [14], [16], [19], [20], [22], [24], [30], [37], [34], [43], [54], [63], [64], [71], [76], [77], [79], [83], and top-k query processing [74], [43], [54], [16], [19], [6], [28], [38], [41], [73]. Techniques for performance speed up will be discussed, including indexing [14], [18], [22], [37], [58], materialized views [46], and data source selections [76], [83]. We will discuss evaluation framework for keyword search engines, the INEX (INitiative for the Evaluation of XML Retrieval) benchmark for XML keyword search [26] and a formal evaluation approach using a set of axioms that captures broad intuitions [47].

Compared to the related tutorial offered earlier, this tutorial will focus on the latest research advances in this area [80], [60], [4], [72], [39], [66], [10], [3], [44], [36], [62], [51], [61], [35], [9], [12], [40]. These recent developments not only provide more efficient and scalable solutions for keyword search on databases, but also open up several new research topics that are worth further investigation, Techniques that refine user queries or help users issue queries will be discussed, such as query cleaning [59], [53], semantic-driven approximate match [80], query auto-completion [9], [36], [35], and query expansion [62]. We will discuss techniques that help users to judge result relevance and analyze the results, including result snippets [25], [49], result clustering [23], [33], [77], [85], [48], result comparison [51], and personalization [66]. Techniques on authority flow based ranking [75], [7] and domain-specific search [17] will also be discussed.

We will also discuss variations of keyword search [71], [79], query form generation for database access [69], and the combination of search and form-based access [12]. Besides presenting the techniques of supporting keyword search on relational databases, graph-structured data and XML data, we will also discuss how to support keyword search on other data models, such as data streams [57], [56], workflows [50], [65], spatial and multimedia databases [15], [84], uncertain data [40], and relationship among them.

We will introduce research challenges and the state-of-the-art of these problems, and discuss their relationships, in order to provide the audience with a big picture of supporting keyword-based search and exploration on structured data.

We will identify and analyze opportunities for future research to advance the field. For instance, how should we support diverse and heterogeneous data models? How should we strike a good balance between the expressiveness and the simplicity of the query language? What are the unique opportunities to analyze the results of searching databases? How can we effectively combine the information in query logs and in databases to enhance the search quality?

III. About the Presenters

Yi Chen is an Assistant Professor in the Department of Computer Science and Engineering at Arizona State University, USA. She received Ph.D. degree in Computer Science from the University of Pennsylvania in 2005. She is a recipient of an NSF CAREER award and an IBM faculty award. Her current research interests focus on empowering non-expert users to easily access diverse structured data, in particular, searching and optimization in the context of databases, information integration, workflows, and social network (http://www.public.asu.edu/~ychen127/).

Wei Wang is a Senior Lecturer in the School of Computer Science and Engineering at the University of New South Wales, Australia. He received his Ph.D. degree in Computer Science from Hong Kong University of Science and Technology in 2004. His recent research interests are integration of database and information retrieval technologies, similarity search, and spatial-temporal databases (http://www.cse.unsw.edu.au/~weiw/).

Ziyang Liu is a Ph.D. candidate and an SFAz (Science Foundation Arizona) Graduate Fellowship recipient in the Department of Computer Science and Engineering at Arizona State University. He joined Arizona State University in August 2006 and received M.S. degree in Computer Science in May 2008. His current research focuses on keyword search on structured and semi-structured data and workflow management (http://www.public.asu.edu/~zliu41/).

IV. Acknowledgement

Yi Chen is supported by an NSF CAREER award IIS-0845647 and NSF grant IIS-0740129, IIS-0915438 and an IBM faculty award. Wei Wang is supported by ARC Discovery Grants DP0987273 and DP0881779.

References

[1] S. Agrawal, S. Chaudhuri, and G. Das. DBXplorer: A system for keyword-based search over relational databases. In *ICDE*, pages 5–16, 2002.

[2] S. Amer-Yahia and J. Shanmugasundaram. XML full-text search: Challenges and opportunities. In *VLDB*, page 1368, 2005.

[3] A. Baid, I. Rae, A. Doan, and J. Naughton. Toward Industrial-Strength Keyword Search Systems over Relational Data. In *ICDE*, 2010.

[4] A. Baid, I. Rae, J. Li, A. Doan, and J. F. Naughton. Toward Scalable Keyword Search over Relational Data. *PVLDB*, 3(1):140–149, 2010.

[5] Z. Bao, T. W. Ling, B. Chen, and J. Lu. Effective XML Keyword Search with Relevance Oriented Ranking. In *ICDE*, 2009.

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

[7] S. Chakrabarti. Dynamic personalized pagerank in entity-relation graphs. In *WWW*, pages 571–580, 2007.

[8] S. Chaudhuri and G. Das. Keyword querying and ranking in databases. *PVLDB*, 2(2):1658–1659, 2009.

[9] S. Chaudhuri and R. Kaushik. Extending Autocompletion to Tolerate Errors. In *SIGMOD*, 2009.

[10] L. Chen and Y. Papakonstantinou. Supporting Top-K Keyword Search in XML Databases. In *ICDE*, 2010.

[11] Y. Chen, W. W. 0011, Z. Liu, and X. Lin. Keyword search on structured and semi-structured data. In *SIGMOD Conference*, pages 1005–1010, 2009.

[12] E. Chu, A. Baid, X. Chai, A. Doan, and J. Naughton. Combining Keyword Search and Forms for Ad Hoc Querying of Databases. In *SIGMOD*, 2009.

[13] S. Cohen, J. Mamou, Y. Kanza, and Y. Sagiv. XSEarch: A semantic search engine for XML. In *VLDB*, 2003.

[14] B. B. Dalvi, M. Kshirsagar, and S. Sudarshan. Keyword search on external memory data graphs. *PVLDB*, 1(1):1189–1204, 2008.

[15] I. De Felipe, V. Hristidis, and N. Rishe. Keyword search on spatial databases. In *ICDE*, pages 656–665, 2008.

[16] B. Ding, J. X. Yu, S. Wang, L. Qin, X. Zhang, and X. Lin. Finding top-k min-cost connected trees in databases. In *ICDE*, 2007.

[17] F. Farfán, V. Hristidis, A. Ranganathan, and M. Weiner. Xontorank: Ontology-aware search of electronic medical records. In *ICDE*, pages 820–831, 2009.

[18] R. Goldman, N. Shivakumar, S. Venkatasubramanian, and H. Garcia-Molina. Proximity search in databases. In *VLDB*, pages 26–37, 1998.

[19] K. Golenberg, B. Kimelfeld, and Y. Sagiv. Keyword proximity search in complex data graphs. In *SIGMOD*, 2008.

[20] L. Guo, F. Shao, C. Botev, and J. Shanmugasundaram. XRANK: Ranked keyword search over XML documents. In *SIGMOD*, 2003.

[21] A. Y. Halevy, M. J. Franklin, and D. Maier. Principles of dataspace systems. In *PODS*, pages 1–9, 2006.

[22] H. He, H. Wang, J. Yang, and P. S. Yu. Blinks: ranked keyword searches on graphs. In *SIGMOD*, pages 305–316, 2007.

[23] V. Hristidis, N. Koudas, Y. Papakonstantinou, and D. Srivastava. Keyword proximity search in XML trees. *IEEE Transactions on Knowledge and Data Engineering*, 18(4), 2006.

[24] V. Hristidis, Y. Papakonstantinou, and A. Balmin. Keyword proximity search on xml graphs. In *ICDE*, 2003.

[25] Y. Huang, Z. Liu, and Y. Chen. Query biased snippet generation in XML search. In *SIGMOD*, pages 315–326, 2008.

[26] INEX. Initiative for the evaluation of xml retrieval. http://inex.is.informatik.uni-duisburg.de/.

[27] H. V. Jagadish, A. Chapman, A. Elkiss, M. Jayapandian, Y. Li, A. Nandi, and C. Yu. Making database systems usable. In *SIGMOD*, pages 13–24, 2007.

[28] V. Kacholia, S. Pandit, S. Chakrabarti, S. Sudarshan, R. Desai, and H. Karambelkar. Bidirectional expansion for keyword search on graph databases. In *VLDB*, pages 505–516, 2005.

[29] KEYS 2009. The First International Workshop on Keyword Search on Structured Data, 2009.

[30] B. Kimelfeld and Y. Sagiv. Finding and approximating top-k answers in keyword proximity search. In *PODS*, pages 173–182, 2006.

[31] L. Kong, R. Gilleron, and A. Lema. Retrieving Meaningful Relaxed Tightest Fragments for XML Keyword Search. In *EDBT*, 2009.

[32] G. Koutrika, A. Simitsis, and Y. E. Ioannidis. Précis: The essence of a query answer. In *ICDE*, page 69, 2006.

[33] G. Koutrika, Z. M. Zadeh, and H. Garcia-Molina. DataClouds: Summarizing Keyword Search Results over Structured Data. In *EDBT*, 2009.

[34] G. Li, J. Feng, J. Wang, and L. Zhou. An effective and versatile keyword search engine on heterogenous data sources. *PVLDB*, 1(2):1452–1455, 2008.

[35] G. Li, S. Ji, C. Li, and J. Feng. Efficient Type-Ahead Search on Relational Data: A TASTIER Approach. In *SIGMOD*, 2009.

[36] G. Li, S. Ji, C. Li, J. Feng, and J. Wang. Efficient Fuzzy Type-Ahead Search in TASTIER (demo). In *ICDE*, 2010.

[37] G. Li, B. C. Ooi, J. Feng, J. Wang, and L. Zhou. EASE: an effective 3-in-1 keyword search method for unstructured, semi-structured and structured data. In *SIGMOD*, 2008.

[38] G. Li, X. Zhou, J. Feng, and J. Wang. Progressive Top-k Keyword Search in Relational Database. In *ICDE*, 2009.

[39] J. Li, C. Liu, R. Zhou, and W. Wang. Suggestion of Promising Result Types for XML Keyword Search. In *EDBT*, 2010.

[40] J. Li, C. Liu, R. Zhou, and W. Wang. Top-k Keyword Search over Probabilistic XML Data. In *ICDE*, 2011.

[41] W.-S. Li, K. S. Candan, Q. Vu, and D. Agrawal. Retrieving and organizing web pages by "information unit". In *WWW*, pages 230–244, 2001.

[42] Y. Li, C. Yu, and H. V. Jagadish. Schema-free XQuery. In *VLDB*, 2004.

[43] F. Liu, C. Yu, W. Meng, and A. Chowdhury. Effective keyword search in relational databases. In *SIGMOD*, pages 563–574, 2006.

[44] Z. Liu, Y. Cai, and Y. Chen. TargetSearch: A Ranking Friendly XML Keyword Search Engine (demo). In *ICDE*, 2010.

[45] Z. Liu and Y. Chen. Identifying meaningful return information for xml keyword search. In *SIGMOD*, 2007.

[46] Z. Liu and Y. Chen. Answering keyword queries on XML using materialized views. In *ICDE*, 2008.

[47] Z. Liu and Y. Chen. Reasoning and identifying relevant matches for xml keyword search. *PVLDB*, 1(1):921–932, 2008.

[48] Z. Liu and Y. Chen. Return specification inference and result clustering for keyword search on xml. *ACM Trans. Database Syst.*, 35(2), 2010.

[49] Z. Liu, Y. Huang, and Y. Chen. Improving xml search by generating and utilizing informative result snippets. *ACM Trans. Database Syst.*, 35(3), 2010.

[50] Z. Liu, Q. Shao, and Y. Chen. Searching workflows with hierarchical views. *PVLDB*, 3(1):918–927, 2010.

[51] Z. Liu, P. Sun, and Y. Chen. Structured Search Result Differentiation. In *VLDB*, 2009.

[52] Z. Liu, J. Walker, and Y. Chen. XSeek: A semantic XML search engine using keywords. In *VLDB*, 2007.

[53] Y. Lu, W. Wang, J. Li, and C. Liu. XClean: Providing Valid Spelling Suggestions for XML Keyword Queries. In *ICDE*, 2011.

[54] Y. Luo, X. Lin, W. Wang, and X. Zhou. SPARK: Top-k keyword query in relational databases. In *SIGMOD*, pages 115–126, 2007.

[55] Y. Luo, W. Wang, and X. Lin. Spark: A keyword search engine on relational databases. In *ICDE*, pages 1552–1555, 2008.

[56] A. Markowetz, Y. Yang, and D. Papadias. Keyword search on relational data streams. In *SIGMOD '07: Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, pages 605–616, New York, NY, USA, 2007. ACM.

[57] A. Markowetz, Y. Yang, and D. Papadias. Keyword search over relational tables and streams. *ACM Trans. Database Syst.*, 34(3), 2009.

[58] A. Markowetz, Y. Yang, and D. Papadias. Reachability Indexes for Relational Keyword Search. In *ICDE*, 2009.

[59] K. Q. Pu and X. Yu. Keyword query cleaning. *PVLDB*, 1(1):909–920, 2008.

[60] L. Qin, J. Yu, and L. Chang. Ten Thousand SQLs: Parallel Keyword Queries Computing. *PVLDB*, 3(1):58–69, 2010.

[61] L. Qin, J. X. Yu, and L. Chang. Keyword Search in Databases: The Power of RDBMS. In *SIGMOD*, 2009.

[62] N. Sarkas, N. Bansal, G. Das, , and N. Koudas. Measure-driven Keyword-Query Expansion. In *VLDB*, 2009.

[63] M. Sayyadian, H. LeKhac, A. Doan, and L. Gravano. Efficient keyword search across heterogeneous relational databases. In *ICDE*, pages 346–355, 2007.

[64] F. Shao, L. Guo, and C. Botev. Efficient Keyword Search over Virtual XML Views. In *VLDB*, 2007.

[65] Q. Shao, P. Sun, and Y. Chen. WISE: a workflow information search engine. In *ICDE*, 2009.

[66] K. Stefanidis, M. Drosou, and E. Pitoura. PerK: Personalized Keyword Search in Relational Databases through Preferences. In *EDBT*, 2010.

[67] C. Sun, C.-Y. Chan, and A. Goenka. Multiway SLCA-based keyword search in XML data. In *WWW*, 2007.

[68] Databases and IR: Perspectives of a SQL guy. NSF Information and Data Management PI Workshop, 2003.

[69] P. P. Talukdar, M. Jacob, M. S. Mehmood, K. Crammer, Z. G. Ives, F. Pereira, and S. Guha. Learning to create data-integrating queries. *PVLDB*, 1(1):785–796, 2008.

[70] Y. Tao and J. X. Yu. Finding Frequent Co-occurring Terms in Relational Keyword Search. In *EDBT*, 2009.

[71] S. Tata and G. M. Lohman. SQAK: doing more with keywords. In *SIGMOD*, pages 889–902, 2008.

[72] A. Termehchy and M. Winslett. Keyword Search for Data-Centric XML Collections with Long Text Fields. In *EDBT*, 2010.

[73] T. Tran, S. Rudolph, P. Cimiano, and H. Wang. Top-k Exploration of Query Candidates for Efficient Keyword Search on Graph-Shaped (RDF) Data. In *ICDE*, 2009.

[74] L. G. Vagelis Hristidis and Y. Papakonstantinou. Efficient ir-style keyword search over relational databases. In *VLDB*, 2003.

[75] R. Varadarajan, V. Hristidis, and L. Raschid. Explaining and reformulating authority flow queries. In *ICDE*, pages 883–892. IEEE, 2008.

[76] Q. H. Vu, B. C. Ooi, D. Papadias, and A. K. H. Tung. A graph method for keyword-based selection of the top-k databases. In *SIGMOD*, 2008.

[77] S. Wang, Z. Peng, J. Zhang, L. Qin, S. Wang, J. X. Yu, and B. Ding. NUITS: A novel user interface for efficient keyword search over databases. In *VLDB*, pages 1143–1146, 2006.

[78] G. Weikum. DB&IR: both sides now. In *SIGMOD*, pages 25–30, 2007.

[79] P. Wu, Y. Sismanis, and B. Reinwald. Towards keyword-driven analytical processing. In *SIGMOD*, pages 617–628, 2007.

[80] D. Xin, Y. He, and V. Ganti. Keyword++: A Framework to Improve Keyword Search Over Entity Databases. *PVLDB*, 3(1):711–722, 2010.

[81] Y. Xu and Y. Papakonstantinou. Efficient keyword search for smallest LCAs in XML databases. In *SIGMOD*, 2005.

[82] Y. Xu and Y. Papakonstantinou. Efficient LCA based Keyword Search in XML Data. In *EDBT*, 2008.

[83] B. Yu, G. Li, K. R. Sollins, and A. K. H. Tung. Effective keyword-based selection of relational databases. In *SIGMOD*, pages 139–150, 2007.
[84] D. Zhang, Y. M. Chee, A. Mondal, A. Tung, and M. Kitsuregawa. Keyword Search in Spatial Databases: Towards Searching by Document. In *ICDE*, 2009.
[85] B. Zhou and J. Pei. Answering Aggregate Keyword Queries on Relational Databases Using Minimal Group-bys. In *EDBT*, 2009.