Information Integration: Challenges and Approaches

Sharma Chakravarthy

IT Laboratory and Computer Science and Engineering Department
The University of Texas at Arlington, Arlington, TX 76019-0019, USA
***email*:sharma@cse.uta.edu, *url*: http://itlab.uta.edu/sharma**

Abstract

The problem of information integration is discussed in the context of answering a query over the web. Querying the web requires that information from different web and other sources be intelligently combined to generate all or top-k answers. A number if issues from query specification to extraction and integration of partial results to ranking of results needs to be addressed for this to happen seamlessly.

1 Audience

Practitioners and professionals, requiring up-to-date information on latest trends, in newer forms of accessing information on the Web (in addition to search and meta search) will benefit from this tutorial. The presenter has been working, for a while, on information integration, query-by-keywords, and ranking results in the context of Web queries. Although internet search itself has been around for a while and is used by general populace as well as by technical users, querying the Web is still in its infancy and is limited to specific domains and applications (e.g., airline reservation). In contrast, querying a structured database has been around for several decades and query answering as well query optimization has advanced to a significant stage.

The challenge now is whether the work on querying a structured database can be redirected meaningfully towards querying the web? This tutorial will address this problem in detail bringing out the current state of the art as well as advances that may help in addressing the problem.

Unlike search, querying the internet requires intelligent integration of information from multiple web sources to construct meaningful answers. A number of new issues, such as how to pose a query (by non-technical users), how to determine sources for answering a query (known as source discovery), how to deal with lack of schema (schema discovery), data extraction from web sources, web query optimization, integration/combining data from multiple sources, and ranking of results, need to be addressed in order to solve the problem of information integration. Several other issues, such as coverage of information, trust/confidence of information, where to access information if there are multiple sources, are also important.

We will present several approaches to information integration from different perspectives that have been proposed in the literature. We will also present applications that can benefit querying the web that requires information integration. Practitioners will benefit from the practical nature of the topics and find the solutions presented applicable to problems they have encountered. Researchers will benefit from the issues that need to be addressed in one of the hot areas currently being revolutionized by increasing amount of information available over the internet.

2 Description

In this tutorial, we introduce the difference between search, meta search, and information integration. We also introduce the differences between query processing over a database (where the schema and statistics are available) and query processing over the internet. We survey a number of techniques and systems that have attempted information integration in specific contexts. We briefly survey Havasu [9], Ariadne [10], MetaQuerier [2], Information Manifold [12], Informaster [3], Tukwila [8], and others. More recent work on information integration which focuses on analytics of different data formats will be covered [6, 1, 5]. We then introduce *InfoMosaic*, its framework and architecture that is being researched by the presenter at UTA [17, 16, 18, 19, 15, 14].

We then introduce a general framework and an architecture (***InfoMosaic***) to highlight the subproblems involved in processing an arbitrary query over the internet. This will bring out a number of new problems, their complexity, and where they stand currently in terms of solutions. This set of problems

15th International Conference on Management of Data COMAD 2009, Mysore, India, December 9–12, 2009

include: query specification, determining sources for answering a query (known as source discovery), dealing with lack of schema (schema discovery), query optimization, data extraction from web sources, integration/combining data from multiple sources, and ranking of results, to name a few. Finally, we discuss a few of the above problems in detail and elaborate on the approaches being used in the literature [13, 11, 7]. Specifically, we will address query specification, query optimization, and ranking of results in detail.

To summarize, we first overview earlier work on information integration applicable to limited/special contexts. We then present the general purpose problem along with details of sub-problems that need to be solved in order to accomplish true information integration. We then elaborate on a few problems and indicate how they are being addressed in the literature. We will also contrast the differences and similarities between traditional query processing on known schema and data sets with query processing on the web.

References

[1] P. A. Bernstein and L. M. Haas. Information integration in the enterprise. *Commun. ACM*, 51(9):72–79, 2008.

[2] K. C.-C. Chang, B. He, and Z. Zhang. Toward Large Scale Integration: Building a MetaQuerier over Databases on the Web. In *Conference on Innovations in Database Research (CIDR)*, pages 44–55, 2005.

[3] O. M. Duschka and M. R. Genesereth. Infomaster: An Information Integration Tool. In *International Conference on Intelligent Information Integration*, 1997.

[4] A. Gal. Why is schema matching tough and what can we do about it? *SIGMOD Record*, 35(4):2–5, 2006.

[5] L. M. Haas. Beauty and the beast: The theory and practice of information integration. In *ICDT*, pages 28–43, 2007.

[6] L. M. Haas and A. Soffer. New challenges in information integration. In *DaWaK*, pages 1–8, 2009.

[7] H.Wache, T. Vogele, U. Visser, H. Stuckenschmidt, G. Schuster, H. Neumann, and S. Hubner. Ontology-Based Information Integration: A Survey. *International Journal on Artificial Intelligence*, 2001.

[8] Z. G. Ives, D. Florescu, M. Friedman, A. Levy, and D. S. Weld. An Adaptive Query Execution System for Data Integration. In *SIGMOD Conference*, 1999.

[9] S. Kambhampati, U. Nambiar, Z. Nie, and S. Vaddi. Havasu: A Multi-Objective, Adaptive Query Processing Framework for Web Data Integration. Technical report, Arizona State University, 2002.

[10] C. A. Knoblock, S. Minton, J. L. Ambite, N. Ashish, I. Muslea, A. Philpot, and S. Tejada. The Ariadne Approach to Web-Based Information Integration. *International Journal on Cooperative Information Systems*, 10(1-2):145–169, 2001.

[11] M. Lenzerini. Data Integration: A Theoretical Perspective. In *ACM Symposium on Principles of Database Systems (PODS)*, pages 233–246, 2002.

[12] A. Y. Levy. Information Manifold Approach to Data Integration. *IEEE Intelligent Systems*, pages 1312–1316, 1998.

[13] A. P. Sheth and J. A. Larson. Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases. *ACM Computing Surveys*, 22(2):183–236, 1990.

[14] A. Telang and S. Chakravarthy. Information Integration across Heterogeneous Domains: Current Scenario, Challenges and the InfoMosaic Approach. Technical report, University of Texas at Arlington, 2007.

[15] A. Telang, S. Chakravarthy, and Y. Huang. Information integration across heterogeneous sources: Where do we stand and how to proceed? In *International Conference on Management of Data (COMAD)*, pages 186–197, 2008.

[16] A. Telang, S. Chakravarthy, and C. Li. Querying for information integration: How to go from an imprecise intent to a precise query? In *International Conference on Management of Data (COMAD)*, pages 245–248, 2008.

[17] A. Telang, S. Chakravarthy, and C. Li. Query-By-Keywords (QBK): Query Formulation Using Semantics and Feedback. In *International Conference on Conceptual Modeling (ER)*, 2009.

[18] A. Telang, C. Li, and S. Chakravarthy. One Size Does Not Fit All: Towards User- & Query-Dependent Ranking For Web Databases. Technical Report 6, University of Texas at Arlington, 2009.

[19] A. Telang, R. Mishra, and S. Chakravarthy. Ranking Issues for Information Integration. In 257-260, editor, *International Conference on Data Engineering (ICDE) Workshop (DBRank)*, 2007.