

Accessing the Web: From Search to Integration*

Kevin Chen-Chuan Chang
Computer Science Department
University of Illinois at Urbana-Champaign
kcchang@cs.uiuc.edu

Junghoo Cho
Computer Science Department
University of California, Los Angeles
cho@cs.ucla.edu

1. OVERVIEW

We have witnessed the rapid growth of the World Wide Web—The Web has not only “broadened” but also “deepened”: A 1999 survey [1] estimated a total of 800 million pages on the Web at that time. Nowadays, 19.2 billion Web pages as reported by the recent index of Yahoo.com, denoting at least 24 times increase in six years. Further, while this *surface Web* has linked billions of static HTML pages, an equally or even more significant amount of information is “hidden” on the *deep Web*, behind the query forms of searchable databases. A July 2000 survey [2] estimated 96,000 “search sites” and 550 billion content pages in this deep Web. A more recent study [3] in April 2004 estimated 330,000 deep Web sources with over 1.2 million query forms, reflecting a fast 3-7 times increase in 4 years. With the virtually unlimited amount of unstructured and structured information sources, the Web is clearly an important frontier for data management and knowledge discovery.

Accessing information on the Web thus requires not only *search* to locate pages of interests, on the surface Web, but also *integration* to aggregate data from alternative or complementary sources, on the deep Web. While the opportunities are unprecedented, the challenges are also immense: For the surface Web, while search seems to have evolved into a standard technology, its maturity and pervasiveness have invited the attack of spam and the demand of personalization. On the other hand, for the deep Web, while the proliferation of structured sources has promised opportunities for more precise and aggregated access, it has presented new challenges for large scale and dynamic information integration. These issues are essentially related to data management, in a large scale, and thus present novel problems and interesting opportunities for our research community.

*This material is based upon work partially supported by the National Science Foundation under Grant No. IIS-0347993, IIS-0133199, and IIS-0313260. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

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 2006, June 27–29, 2006, Chicago, Illinois, USA.
Copyright 2006 ACM 1-59593-256-9/06/0006 ...\$5.00.

Goals. In this tutorial, we will go over recent advances made in helping users access information on the Web, in order to (1) promote the awareness of the database community for the technical challenges in this area, (2) stress the new spirit and shifted agenda injected by the unique characteristics of the Web information, and (3) engage the database community for open issues and research problems.

Target Audience. The tutorial is intended for researchers who are interested in the new access scenarios and research problems in Web information search and integration, from the perspective of the SIGMOD community, *i.e.*—How is Web search different from querying a database? How does Web integration, with its large scale and open borders, differ from traditional information integration scenarios?

Prerequisite. No specific prerequisite knowledge. The tutorial will start from introducing basic characteristics and observations of Web information access, and the technical details will be self-contained.

2. OUTLINE

Two main topics that we will cover in this tutorial are (1) *keyword-based searches* on the surface Web and (2) *source integration* on the deep Web. Traditionally, users have predominantly used a keyword-based search interface in accessing the information on the surface Web, where they type in a short list of keywords that describe what the users are interested in. While this interface has proven to be powerful in describing the user’s need and identifying relevant textual pages on the surface Web, its inherent limitation and the rapid growth of the deep Web has also made it necessary to develop novel ways to integrate diverse sources that provide rich interfaces on structured data. This tutorial will be organized around these two approaches for Web information access and cover their historical evolution, state of the art development, and open research issues.

2.1 Part I: Searching the Surface Web

While having an elegantly simple interface, it requires tremendous research and engineering efforts to build and operate an effective keyword-based search engine. First of all, search engines have to constantly fight against Web spammers who employ an ever-evolving set of tricks to artificially boost their ranking in the search results. They also need to be able to distinguish the diverse need of a user simply based on a few keywords that the user types in. In the first part of this tutorial, we will review some of the recent advances

made by the Web, database and IR community in this context. Roughly, this part of the tutorial will be organized around the following themes:

1. Basics of Web Search
 - General architecture: Web crawler, indexer, and query engine
 - Basic terms and models: inverted index, vector-space model, TF-IDF and PageRank
 - Taxonomy of Web search
2. Fighting Web Spam: How search engines identify and filter out the pages that artificially boost their ranking in search results
 - Content-based spam detection
 - Link-based spam detection
3. Personalized Search: One global ranking for every user may not be adequate when multiple users have diverse needs. How we can customize search results depending on the interest of a particular user
 - Topic-Sensitive PageRank
 - User preference model
 - Personalized ranking
4. Architecture and Performance Issues: How search engines deal with the enormous data and query load and be able to answer queries quickly with reasonable resources
 - Link analysis optimization
 - Query engine scalability
5. Search Engine Impact and Bias: How the ranking in search results may affect how users access Web pages. How we can model these changes in the user behavior and what are the implications of these changes
 - Web-user behavior and models
 - Page popularity evolution
 - New ranking mechanisms
6. Open Research Challenges

2.2 Part II: Integrating the Deep Web

With the proliferation of databases on the Web, their impacts to Web information access is immense. In the recent years, as many research projects have built up significant efforts in addressing the deep Web, their results have started to flourish. Meanwhile, the search industry has moved rapidly, beyond page-based unstructured text, to the previously unexplored frontier of database-backed structured data—*e.g.*, the many "vertical" search engines that cover structured data in certain domains (*e.g.*, jobs, products), and the very recent debut of the "Google Base" for building a Web-scale structured database across various domains. The second part of the tutorial will bring the awareness of this important new frontier, outlined as follows:

1. The Emergence of the Deep Web
 - Basics of the deep Web
 - Surveys of the deep Web
 - Observations and implications
2. The Systems: Web Integration Applications

- Horizontal: Deep-web search systems
 - Vertical: Domain-based search engines
3. The Problems: Technical Issues and State of the Art
 - Source modeling and characterization
 - Schema matching
 - Query mediation
 - Data collection
 - Wrapper construction
 - System integration
 4. The Solutions: Common Techniques & Unified Insights
 - Common basic approaches
 - Emergence of unified insights
 5. Open Issues and Future Directions

3. ACKNOWLEDGMENTS

We thank David DeWitt and Jennifer Widom for their helpful suggestions and discussions in refining the themes of this tutorial. We are also grateful to Bin He for his help in preparing the initial proposal.

4. ABOUT PRESENTERS

Kevin Chen-Chuan Chang is an Assistant Professor in Computer Science at UIUC, with a Ph.D. in Electrical Engineering in 2001 from Stanford University. His research focuses on large scale information access, with current emphasis on Web information integration (the *MetaQuerier* project) and ranked data retrieval (*AIM*). He has served on PC in major conferences (*e.g.*, SIGMOD, WWW, ICDE) and cochaired several special issues and workshops in Web integration and mining (*e.g.*, WIRI'06, IIWeb'06). He has received NSF CAREER Award in 2002, NCSA Faculty Fellow Award in 2003, IBM Faculty Awards in 2004 and 2005, and the Incomplete List of Excellent Teachers at UIUC in 2001 and 2004.

Junghoo Cho is an Assistant Professor in the Department of Computer Science at UCLA, with a Ph.D. in Computer Science in 2001 from Stanford University. In the past ten years, he has been conducting research in the study of the evolution, management and retrieval of information on the World-Wide Web. He is an editor of IEEE Internet Computing and serves on program committees of top international conferences, including SIGMOD, VLDB and WWW. He is a recipient of the NSF CAREER Award and IBM Faculty Award.

5. REFERENCES

- [1] Steve Lawrence and C. Lee Giles. Accessibility of information on the web. *Nature*, 400(6740):107–109, 1999.
- [2] BrightPlanet.com. The deep web: Surfacing hidden value. Accessible at <http://brightplanet.com>, July 2000.
- [3] Kevin Chen-Chuan Chang, Bin He, Chengkai Li, Mitesh Patel, and Zhen Zhang. Structured databases on the web: Observations and implications. *SIGMOD Record*, 33(3):61–70, September 2004.