Towards Next Generation Web Information Retrieval

Wei-Ying Ma, Hongjiang Zhang, and Hsiao-Wuen Hon

Microsoft Research Asia

Abstract. Today search engines have become one of the most critical applications on the Web, driving many important online businesses that connect people to information. As the Web continues to grow its size with a variety of new data and penetrate into every aspect of people's life, the need for developing a more intelligent search engine is increasing. In this talk, we will briefly review the current status of search engines, and then present some of our recent works on building next generation web search technologies. Specifically, we will talk about how to extract data records from web pages using vision-based approach, and introduce new research opportunities in exploring the complementary properties between the surface Web and the deep Web to mutually facilitate the processes of web information extraction and deep web crawling. We will also present a search prototype that data-mines deep web structure to enable one-stop search of multiple online web databases.

In contrast with current web search that is essentially document-level ranking and retrieval, an old paradigm in IR for more than 25 years, we will introduce our works in building a new paradigm called object-level web search that aims to automatically discover sub-topics (or taxonomy) for any given query and put retrieved web documents into a meaningful organization. We are developing techniques to provide object-level ranking, trend analysis, and business intelligence when the search is intended to find web objects such as people, papers, conferences, and interest groups.

We will also talk about vertical search opportunities in some emerging new areas such as mobile search and media search. In addition to providing information adaptation on mobile devices, we believe location-based and context-aware search is going to be important for mobile search. We also think that by bridging physical world search to digital world search, many new user scenarios that do not yet exist on desktop search can potentially make a huge impact on the mobile Internet. For media search, we will present those new opportunities in analyzing the multi-typed interrelationship between media objects and other content such as text, hyperlinks, deep web structure, and user interactions for better semantic understanding and indexing of media objects. We will also discuss our goal of continually advancing web search to next level by applying data mining, machine learning, and knowledge discovery techniques into the process of information analysis, organization, retrieval, and visualization.

X. Zhou et al. (Eds.): WISE 2004, LNCS 3306, p. 17, 2004.