CS−Engine: A Cross−Search Engine for Searching Heterogeneous Multiple Databases

Min Song and Il−Yeol Song
College of Information Science and Technology
Drexel University
Philadelphia, PA 19104
(Min.song, songiy)@Drexel.edu

As the popularity and complexity of Internet search engines increase, the design, development and maintenance of large, complex web−based Information Retrieval (WIR) systems are becoming more complex and difficult. The difficulty designing a WIR system is compounded by information overload triggered from various different information sources. From a standpoint of the search engine users, it is more usable for the WIR to provide a single search point to multiple databases. In this paper, we present the design and implementation of a cross−search component for CS−Engine (Cross−Search Engine), a WIR system. The CS−Engine allows the user to search heterogeneous, multiple databases. That is, the CS−Engine is differentiated from other search engines in that it can allow the user to search both public domain web data and proprietary databases of a company. The CS−Engine is also different from meta−search engines because the CS−Engine does not need to trigger other search engines and translates a query to search other engines. Cross−search capability provided by CS−Engine alleviates the users' inconvenience of switching databases and looking up the Internet to satisfy their information needs. The CS−Engine aims to serve as an "Information Mall" to search multiple databases ranging from publicly available database such as web data to subscription−based private databases such as Dialog databases. In the section of the lesson learned, we discuss technical issues as well as organizational issues encountered during the development phase of the system.

A typical Web IR system consists of three components: 1) crawler, 2) index, and 3) search component [1]. CS−Engine adopts a generic architecture proposed by Brian and Page [1]. Figure 1 illustrates a simplified overview of CS−Engine architecture. Although this paper aims to focus on the search component, the other components are mentioned briefly to help the reader understand the architecture of CS−Engine.

Figure 1. The Overall System Architecture. The component surrounded by dashed lines represent CS−Engine.

[Brian and Page, 1998] Brin, Sergey and Page, Lawrence, *The Anatomy of a Large−Scale Hypertextual Web Search Engine* 7th International World Wide Web Conference, Brisbane, Australia pages 14−−18 April 1998.