WebBase and the Stanford InterLib Project
(Extended Abstract)

Junghoo Cho    Hector Garcia-Molina
Department of Computer Science
Stanford, CA 94305
{cho, hector}@cs.stanford.edu

At Stanford, we have been developing base technologies that are required to overcome some of the most critical barriers to effective digital libraries. One of these barriers is the heterogeneity of information and services. Another impediment is the lack of powerful filtering mechanisms that let users find truly valuable information. Third, the continuous access to information is restricted by the unavailability of library interfaces and tools that effectively operate on portable devices. A fourth barrier is the lack of a solid economic infrastructure that encourages providers to make information available, and gives users privacy guarantees. A fifth barrier is the loss of digital information, due to obsolescence of equipment or software.

The Stanford InterLib Technologies Project is developing mechanisms for surmounting these barriers. For example, we have developed a simple interoperability protocol, SDLIP, and have shown how it can be used to access collections at many libraries. We are developing filtering mechanisms that can find information based on its value, as opposed to simply its textual similarity to some query terms [16, 5]. We have build a PowerBrowser that makes it possible for users to access digital information in a convenient way from mobile, hand-held devices [4, 2, 3]. We have also developed an archival repository that makes it possible for autonomous organizations to cooperate in preserving digital documents [13, 12, 14].

The Stanford WebBase is a system which stores a significant portion of the Web in a central repository [15, 1]. By caching the full Web, the user can conveniently access the Web, even when she has intermittent or limited access to the global Web. Also based on this central data, researchers can study the Web data with ease, and can experiment with new search and data mining techniques.

However, it is challenging to collect and manage large amounts of rapidly changing Web data. The Web is estimated to have at least several billion pages and a big portion of the Web changes very often, sometimes more than once a day. Thus the system should be able to download and manage a huge amount of data in an efficient way. This task is often done by a program, called a Web crawler, which needs to address many important challenges including the following:

- **Page Selection**: The size of the Web is huge, so most crawlers can download only a subset of the entire Web. In this context, it is important that a crawler downloads "important" or "relevant" pages first, so that the quality of the downloaded collection is maximized. Clearly, the definition of an "important page" may differ between applications. At Stanford, we studied various importance metrics that have been proposed in literature and are being used by various search engines. Then we designed algorithms that identify important pages early for each importance metric [10, 11].

- **Parallelization**: Since the crawler has to download hundreds of millions pages in a short period of time, the crawler often needs to be parallelized on multiple machines. Clearly, these parallel crawlers should be properly coordinated, so that different crawlers do not visit the same web site multiple times. We studied how we can effectively parallelize a crawling process, so that the coordination overhead is minimized while the download rate is maximized [6, 8].

- **Synchronization**: Once the crawler downloaded pages and stored them locally, the crawler has to refresh the pages periodically, because pages are constantly updated. Given that Web pages change at very different rates, we may make this refresh process much more effective, by adjusting page revisit frequency based on page change frequency. We studied how we can maximize the freshness of the downloaded pages when pages change at different rates [9, 7].

- **Controllability**: When the crawler downloads a page from another Web site, it essentially uses resources belonging to other organizations. For example, the site first needs to read the page from a disk, thus consuming its I/O bandwidth, and the page needs to be
transferred through a shared network. Therefore, it is important that the crawler is highly configurable and controllable so that we adjust its behavior depending on the requirement of a site. The Stanford WebBase crawler was designed from the onset with this goal in mind and can accommodate varying requirements of different Web sites.

Our project is a component of the InterLib collaboration between University of California at Berkeley, the University of California at Santa Barbara, and Stanford University. The combined InterLib technologies will be demonstrated on the emerging California Digital Library (CDL), and on a testbed developed by the San Diego Supercomputer Center.

References


