Information Integration in Digital Library Environment

Aug. 10-11, 2000

Sang Ho Lee, Associate Professor
School of Computing, Soongsil University, Seoul, Korea
Email: shlee@computing.soongsil.ac.kr

As the number of digital libraries that have been developed and are under operation increases recently, an attempt that merges existing digital libraries in terms of functions and services is becoming important. The integration of existing digital libraries provides users a spectrum of services, and offers a uniform interface by which users can access multiple information sources.

The objective of this research is to develop information integration techniques in digital library environments, where each information source is unique in terms of data formats, data types, data contents and so on. In particular, technical issues in the integration of Web search engines in a mediator approach shall be explored. Integration technique includes processes that enable us to extract data, to process them internally, to present them to users in response to users' queries and to store them for later usage. below are two research items that I would like to explore toward information integration.

(1) New Web search techniques: There are many Web search engines in the world, and the importance of them is well recognized. Existing Web search engines (for example, Lycos, Altavista, Infoseek, etc.) are based on the theory of Information Retrieval, and in particular rank retrieved documents on the basis of query/document similarities, which have been studied over the last thirty years in the field of Information Retrieval. Users start to realize that the theory does not work out well as expected. For instance, the search engines often return too many unrelated documents and outdated documents to users. A number of new technical attempts (for example, www.goto.com, www.google.com, www.askjeeves.com, and www.cupernic.com) have been

presented and are in operation. The objective is to develop a new Web search technique that overcomes the problems of current search engines. Searching is by nature dependent on types of data being searched. One approach under investigation is focused on a meta-search, in which information is categorized in terms of contents so that search techniques suitable for each category could be developed.

(2) Intelligent wrapper technology: Development of wrappers in Internet databases has been around for several years. A wrapper is used to extract, combine, and reconcile information for several independent information sources. It is part of a mediator-based architecture, which is a popular system architecture for information integration. The wrapper should be created and maintained for each of external information sources in a semi-automated fashion, and should be adaptable for dynamic changes of HTML pages, at least in a certain degree. The wrapper technology is a key ingredient to the integration of HTML documents. Various kinds of wrappers have been developed in universities and research institutes as part of development effort of Web databases. There are also several approaches to build wrappers, including formal language theory based, machine learning based, heuristic based, and extended context-free grammar based, etc.

Biographical sketch: Prof. Sang Ho Lee is an associate professor in School of Computing at Soongsil University, Seoul, Korea. His research interests include benchmark and functional validation of database systems, Web databases, and digital library. In 1990 – 1992, he worked as a senior research member in Electronics and Telecommunications Research Institute (ETRI), which is a government-sponsored, telecommunication-oriented research institute in Deajeon, Korea. From August 1999 to July 2000, He was with George Mason University, Virginia, as a visiting professor. He received his MS and Ph.D degree in Computer Science from Northwestern University, Illinois, in 1986 and 1989, respectively. He received his BS in Computer Engineering from Seoul National University, Korea, in 1984.