Federated Searching across a Language Boundary for an Educational Use of Digital Libraries

Sung Hyon Myaeng
Division of Information & Communication
Chungnam National University
Taejon, Korea
shmyaeng@cs.cnu.ac.kr

1. Introduction

Our group involving several faculty members in Chugnam National University has conducted research on digital libraries for the past few years. In 1996, we started up with a practical goal of building a prototype for general-purpose digital libraries, focusing on the retrieval engine, the Z39.50 protocol, and the user interface [1, 2]. We then attempted to expand the notion of conventional digital libraries being a searchable repository of documents to an environment where users can create new documents using existing ones. Our notion of virtual documents plays an essential role. Along this line of research, we developed a system architecture and implemented the first prototype [3].

Another area of research related to the US-Korea joint workshop on digital libraries is cross-language information retrieval. We devised a practical yet effective algorithm for Korean-English cross-language retrieval. A Korean query is translated into an English version based on a bilingual dictionary, a most available type of resource for cross-language IR. The ambiguity problem is handled by exploiting the target language corpus that should be always available as the target of retrieval [4].

Given our background, wee would like to pursue the following goals in a US-Korea joint project on digital libraries.

- To develop techniques for federated searching in multi-lingual, heterogeneous digital library (DL) environment
- To provide a technical basis for an information gathering environment compliant with Open Archives Initiative (OAI) using the aforementioned techniques
- To advance our on-going research on "DL as a dynamic knowledge space" in such a way that personal DLs can be created and used for educational purposes in conjunction with OAI

2. Proposed Project Description

The main thrust of this project is to develop core technology for providing an educational environment where research and educational materials can be shared through internationally interoperable digital libraries. By developing high-performance techniques that allow for retrieving information from

distributed, multi-lingual information repositories, we make it possible for users to efficiently gather information from diverse sources using the native language.

The federated searching techniques will be applied not only to existing well-established digital library resources and Internet materials, but also other repositories that contain grey literature. As such, we will attempt to support and extend the Open Archives Initiatives by collaborating with those researchers heavily involved in the effort.

The federated searching techniques will include two major features. First, we will tackle the problem of collecting, organizing, and delivering high-quality information from multi-lingual, distributed sources using a native language. Second, we will provide a solution to the problem of federating diverse resources and retrieval engines including but not limited to Web-based, OAI-compliant, and Z39.50-compliant digital libraries.

Our proposed research effort will extend our on-going research on "virtual document framework" that facilitates dynamic creation of new multimedia documents using various links and existing materials in distributed locations. By adding the new federated searching functions to be developed in this project, we expect our MIRAGE III system [3] will become much better equipped and significantly enriched for an active learning environment.

3. Biographical Information

Sung Hyon Myaeng is currently a professor at Chungnam National University, Korea. Prior to returning to Korea in 1994, he taught in the School of Information Studies, Syracuse University, for six years, where his tenure was granted. While in Syracuse University, he was a co-PI for the DR-LINK project in the TIPSTER program sponsored by DARPA, where the goal was to develop an information retrieval (IR) system that exploits various semantic features automatically extracted from text and linguistic resources.

His research has centered around information retrieval since his graduate studies. While active in sub-areas of IR such as cross-language IR, summarization, categorization, and distributed IR, he has led several digital library projects in Korea, from a large-scale design at the national level to an implementation project. The MIRAGE system developed by his digital library research group in 1995 is considered the first digital library research prototype developed in Korea, which has been partly commercialized. This prototype has evolved into an architecture that extends the traditional notion of digital libraries into a “dynamic knowledge space” where users can compose, efficiently store, and retrieve new multimedia digital objects from existing distributed resources. The major application of this digital library architecture is likely to be in education and knowledge management in corporate settings.

He is an Associate Editor for ACM Transactions on Asian Language Information Processing and a member of editorial boards for several journals including Information Processing and Management. He has served on program committees for ACM SIGIR for many years as well as other international

conferences on information retrieval, natural language processing, and digital libraries. He started the IRAL (Information Retrieval with Asian Language) Workshop with Ishikawa, whose 5th one will be held in Hong Kong this year.

References

[1] Myaeng, S. H. (1998). "R&D for a Nationwide General-Purpose System" Communications of the ACM, 41 (4), April, 83-85.

[2] Myaeng, S. H. (1996). "MIRAGE: A Prototype for a Multimedia Information Retrieval and Gathering Environment." Proc. of International Conference on Digital Libraries and Information Services for the 21st Century, Seoul, September 10-13.

[3] Myaeng, S. H. et al. (2000). A Digital Library System for Easy Creation & Manipulation of New Documents over Existing Resources," RIAO 2000, Paris, France.

[4] Jang, M.G, Myaeng, S. H., Park, Se Young (1999). "Using Mutual Information to Resolve Query Translation Ambiguities and Query Term Weighting", Proc. of The 37th Annual Meeting of the Association for computational Linguistics (ACL'99), Washington D.C., USA, June.