NanoPort: An Example for Building Knowledge Portals for Scientific Domains

Jialun Qin, Zan Huang, Yilu Zhou, Michael Chau, Chunju Tseng, Alan Yip, T. Gavin Ng,
Fei Guo, Zhi-Kai Chen, Hsinchun Chen
Department of Management Information Systems
The University of Arizona
Tucson, Arizona 85721, USA
qin@u.arizona.edu

Abstract

We describe the NanoPort (www.nanoport.org) system to demonstrate a general framework of building domain-specific knowledge portals. These portals consolidate diverse information resources and provide rich functionalities to support effective information retrieval and knowledge discovery.

Introduction

With increasing academic and research contents available online, the Web has become the largest information repository ever for most scientific domains. However, it has become increasingly difficult to search for high-quality domain-specific information. The resulting information overload problem calls for domain-specific knowledge portals that provide high-quality collection and integrated retrieval and knowledge discovery functionalities for scientific domains. We describe NanoPort, a domain-specific knowledge portal for nanoscale science and engineering (NSSE) field to demonstrate our framework of building such portals.

The NanoPort System

NanoPort integrates and applies several information searching and analysis techniques in the NSSE domain. We describe three major components of the system:

Content Collection Building. We provided both vertical searching and meta-searching features in NanoPort to build a comprehensive NSSE-related information repository: (1) *vertical searching*: we developed a new crawling technique called meta search enhanced global crawling which traverses the Web in a global search manner. The technique keeps expanding starting points by dynamically incorporating relevant meta-search results from other search engines; (2) *meta-searching*: NanoPort also connects to several carefully selected online databases (e.g., MedLine, MatWeb, Molecular Expression, ScienceDirect, and US Patent Database) and journals (e.g. Science, MIT Technology Review, and PNAS).

Retrieval Functionalities. NanoPort integrates several functionalities to support the retrieval process, including: (1) *keyword suggestion*: implemented based on important phrases and relations between them identified by a noun-phrasing tool [2] and a co-occurrence-based automatic thesaurus building tool [1]; (2) *document categorization*: the search results are organized based on key phrases appeared in the documents to provides a content overview of the entire result set; (3) *document summarization*: a summarization tool, AI Summarizer, is embedded into NanoPort to provide document-level sentence-based summarization; and (4) *document visualization*: two self-organizing map-based visualization tools have been employed to generate topic and document maps based on the returned results, a jigsaw-puzzle topic map and a GIS-like document map.

Knowledge Discovery. As a first step of knowledge discovery from text, we analyzed a collection of about 77,000 NSSE-related U.S. patents. Our current efforts include three types of analyses: basic performance evaluation, content map visualization, and citation network analysis. With appropriate text preprocessing, these analyses can also be applied to other types of digital documents.

Acknowledgement

The project is partly supported by the NSF SGER grant "NanoPort: Intelligent Web Searching for Nanoscale Science and Engineering," CTS-0204375, February 2002 - November 2002.

References

[1] Chen, H., and Lynch, K. J. "Automatic construction of networks of concepts characterizing document databases," *IEEE Transactions on Systems, Man and Cybernetics*, 22, 5 (1992), 885-902.

[2] Tolle, K. M., and Chen, H. "Comparing noun phrasing techniques for use with medical digital library tools," *Journal of the American Society of Information Systems*, 51, (2000), 352-370.

JCDL '03 May 1-2, 2003, Houston, Texas.