

# SE-LEGO: Creating Metasearch Engines on Demand

Zonghuan Wu, Vijay Raghavan,  
Chun Du, Komanduru Sai C  
University of Louisiana at Lafayette  
Center for Advan. Computer Studies  
Lafayette, LA 70504  
1-337-482-5243

{zwu,raghavan}@cacs.louisiana.edu

Weiyi Meng, Hai He  
Dept. of Computer Science  
SUNY at Binghamton  
Binghamton, NY 13902  
1-607-777-4311

meng@binghamton.edu

Clement Yu  
Dept. of Computer Science  
University of Illinois at Chicago  
Chicago, IL 60607  
1-312-996-2318

yu@cs.uic.edu

**Categories and Subject Descriptors:** H.3.3 [Information Search and Retrieval]: Information Search and Retrieval

**General Terms:** Experimentation, Design

**Keywords:** Metasearch Engine, Metasearcher, search engine

## Extended Abstract

As a system that provides unified access to multiple existing search systems, a **metasearch engine** can alleviate ordinary users from the formidable task of identifying useful sources and searching them individually. At present, the largest metasearch engines such as ProFusion ([www.profusion.com](http://www.profusion.com)) and SavvySearch ([www.search.com](http://www.search.com)) can connect to about 1,000 search engines. This means that only a small fraction of the information sources on the Web, including both the Surface Web and the Deep Web, are connected, as the number of such sources is estimated to be in the order of hundreds of thousands [1]. Most of these Websites have their own search capabilities and provide search interfaces. Many of these Websites provide high quality information that has been frequently queried by specialists and researchers in particular fields. Present major metasearch engines usually do not connect to these specialized Websites. Currently, building a metasearch engine is an expensive and labor-intensive job that needs diverse expertise. As a result, it is difficult for an ordinary Web user to create a metasearch engine based on the search engines of the user's choice. Some metasearch engine companies (e.g., ProFusion) allow user to build customized metasearch engines, but only search engines in a pre-compiled list can be used because the capability to connect to these search engines need to be established in advance.

In this demonstration, we present a Web-based prototype system called SE-LEGO that is capable of creating metasearch engines on the fly. Only the URLs of the search engines to be used need to be provided to SE-LEGO. As such, SE-LEGO is useful not only for building customized metasearch engines (metasearch engine on demand) for ordinary users but also for building large-scale metasearch engines connecting to tens of thousands or more search engines. The system can be broken into two major modules: (1) Metasearch Engine Generator, and (2) Metasearch Query Processor. The following components are developed to implement the two modules:

- (1) Automatic Search Engine Connection: This component analyzes the source file of the interface of any given search engine and generates a program that can pass queries to the search engine.
- (2) Automatic Search Result Extraction: For any given search engine, this component generates a program to extract the results (e.g., URLs) related to the retrieved documents from the result pages of the search engine.
- (3) Search Engine Discovery: If a user submits a URL that does not have a search engine on the Webpage, this component will crawl Webpages "nearby" to find search engine interfaces.
- (4) Search Engine Representative Collection: For any given search engine, this component generates a representative for the search engine by collecting desired feature information from it.
- (5) Search Engine Selection: For each user query submitted to the metasearch engine, this component selects, based on the representative information of the underlying search engines, a small number of potentially useful search engines to invoke.
- (6) Query Dispatching and Result Merging: This component dispatches queries to selected search engines and merges the results extracted from returned pages into a single ranked list to present to the user.

During the demonstration, we will show how to use SE-LEGO to build a metasearch engine on demand and then conduct metasearching. People in the audience are welcome to provide the URLs of Deep Web and/or Surface Web search engines for the demonstration. We will also demonstrate more advanced features such as Search Engine Selection by using a metasearch engine based on several medical and healthcare search engines such as PubMed (<http://www.ncbi.nih.gov/entrez/query.fcgi>) and WebMD (<http://www.webmd.com>), with their representative collected.

**Acknowledgements:** This work is supported in part by the following grants from NSF (IIS-0208574, IIS-0208434, EIA-9911099), Army Research Office (ARO-2-5-30267), and the IT Initiative of the State of Louisiana to Lafayette.

## REFERENCES

- [1] M. Bergman. The Deep Web: Surfacing the Hidden Value. BrightPlanet White Paper, <http://www.completeplanet.com/Tutorials/DeepWeb/index.asp>, 2000.