SE-Album: A SELEGO Application in Integrated Retrieval From Multiple Online Bio-Informatics Search Systems

Smalheiser, Neil[1], Clement, Yu[1], Torvik, Vetle[*1], Wu, Zonghuan[2], Raghavan, Vijay[2], Qian, Hua[2], Meng, Weiyi[3]
[1]University of Illinois at Chicago, Chicago, IL, USA; [2]University of Louisiana at Lafayette, Lafayette, LA, USA; [3]SUNY Binghamton, Binghamton, NY, USA

Bio-informaticians use the Web in their research more and more frequently. Many rely on general-purpose *search engines*, such as Google or AltaVista. Unfortunately, information found from such search engines typically is not sufficient, since only the *Surface Web* is covered. It is estimated that the *Deep Web* contains hundreds of times more information [1]. As a result, researchers often need to retrieve from a number of bio-informatics Web databases and servers, such as JAX (http://tbase.jax.org/docs/tb.html) and RGD (http://rgd.mcw.edu/), to locate more specialized information. Today, hundreds of such systems are available online, most of which contain rich, high-quality and highly specialized *Deep Web* content that cannot be retrieved from *Surface Web* search engines. Around 200 of them are listed in [2] for illustration.

Obviously, querying against these systems on an individual basis is extremely tedious and time-consuming. As a system that provides unified access to multiple existing search systems, a ***metasearch engine*** **(MSE)** can **relieve users of the formidable task of identifying relevant information sources and searching them separately**. However, major MSEs, as a rule today, do not connect to specialized data sources. For example, ProFusion (http://www.profusion.com), one of the largest MSEs, only connects to about 40 medical search engines, almost all of which only deal with general medical topics, such as diet, pregnancy and health tips.

The major challenge for current MSEs to incorporate many more sources is that such a process calls for special computer expertise, that substantial manual work is needed for each databases, and that it is very difficult to maintain the metasearch engine on a large scale. As a state-of-the-art software technology framework, SELEGO (http://www.selego.com) enables users to create MSEs on the fly. It can be used to build not only customized MSEs for ordinary users, but also large-scale MSEs connecting to tens of thousands of search engines.

In this poster, we introduce SE-Album MSE that draws upon the SELEGO technology. In SE-Album, selected biomedical systems are divided into three groups, based on the query format that a system accepts (1.gene/protein name, 2.accession number, 3.nucleotide/protein sequence). User queries are first reformatted and then sent to individual systems in the corresponding group. After that, meaningful returned results are presented to users for easy browsing and comparison.

For bio-informaticians, SE-Album provides convenient access to specialized Web resources. It is a **novel and convenient approach featuring a common interface that facilitates analysis of the return results**. SELEGO, as its underlying supporting technology, is characterized by highly automated incorporation and maintenance of multiple Web search systems. The quality of SE-Album service is guaranteed through regular automatic system connectivity checking and profile updating. It is also highly scalable in the sense that systems can be conveniently added or removed to meet users' changing requirements.

This work was supported by NSF grants to Drs. Clement Yu and Weiyi Meng, and NIH grants to Dr. Neil Smalheiser.

References

1. M. Bergman. The Deep Web: Surfacing the Hidden Value. BrightPlanet White Paper, 2000.
2. http://nar.oupjournals.org/content/vol30/issue1/