Searching Internet Chemical Information – From Surface Web to Deep Web

LI Xiaoxia*, YUAN Xiaolong, XIA Zhaojie, NIE Fengguang, GUO Li
State Key Lab of Multiphase Complex Systems, Institute of Process Engineering,
Chinese Academy of Sciences, Beijing 100190

The Internet becomes the largest collection and sometimes the only source of chemical information today since the born of World Wide Web around 1995. While enjoying the ever possible convenience in getting information, challenge still exists in developing proper tools for finding scholarly chemical information on Internet because of the dynamic and distributed nature and huge space of the Web. The Web can be classified into surface Web and Deep Web from the point of accessing or index mechanism of general-purpose search engines, where surface Web refers to the portion of the World Wide Web that is indexed by conventional search engines based on hyperlink analysis. The part of the Web that consists of databases and is not reachable this way is called the Deep Web. Although being heavily used daily, what the general-purpose search engines can search are mainly the chemistry surface Web and usually have better recall but lower precision that more refining in searching strategy are needed for users to overcome such limitation. Because of the vast variety of database structures and possible search terms, solutions to search Deep Web are still under development for search engines like Google and other IT explorers, where the chemistry Deep Web has not been covered in these efforts.

Besides the general-purpose search engines, chemistry focused tools are another kind of daily tools for searching the chemistry web. This presentation will overview our efforts in developing chemistry oriented tools for searching both the chemistry surface Web and Deep Web, including a chemistry Web directory, ChIN (http://chin.csdl.ac.cn/) that has been running for 10 years with more than 260,000,000 requests, a prototype chemistry search engine ChemEngine and ChemDB Portal (http://www.chemdb-portal.cn/), a prototype of chemistry Deep Web search engine.

Keywords: Chemistry Search Engine, Chemistry Deep Web, Data extraction

References

Jayant Madhavan, David Ko, Lucja Kot, Vignesh Ganapathy, Alex Rasmussen, Alon Halevy. **Google's Deep-Web Crawl**, *2008 VLDB Endowment, ACM*. http://www.cs.cornell.edu/~lucja/Publications/I03.pdf.

Xiaoxia Li, Li Guo, Xiaolong Yuan, Zhaojie Xia, Fengguang Nie, **Internet Motivated Progress in Chemoinformatics**, *Progress in Chemistry*, **2008**, 20(12):1849

Xiaoxia Li, Xiaolong Yuan, Zhaojie Xia, Fengguang Nie, Wucheng Tang, Li Guo, **A set of discovery tools for Internet chemical information**, *Computers and Applied Chemistry*, 2008, 25(9):1079

Zhaojie Xia, Li Guo, Chunyang Liang, Xiaoxia Li, Zhangyuan Yang, **Focused Crawling for Retrieving Chemical Information**, *Advances in Soft Computing, Innovations in Hybrid Intelligent Systems*, ASC 44, 2007, p. 433-

xxia@home.ipe.ac.cn, Tel. 86-10-62554066 Fax. 86-10-62561822
Thanks for the support of National Science Foundation of China (NSFC20673119, NSFC90612015, NSFC20221603)