ChemXSeer: A Web Search Engine and Repository for e-Chemistry

C. Lee Giles[1,2], Prasenjit Mitra[1,2], Karl Mueller[3], James Z. Wang[1,2], Bingjun Sun[2], Levent Bolelli[2], Xiaonan Lu[2], Ying Liu[1], Isaac Councill[1], William Brower[3], Qingzhao Tan[2], Anuj Jaiswal[1], James Kubicki[4], Barbara Garrison[3], Joel Bandstra[3], Juan Pablo Fernandez Ramirez[1]
[1]Information Sciences and Technology, [2]Computer Science and Engineering, [3]Chemistry, [4]Geosciences
Pennsylvania State University
University Park, PA, 16802 USA
giles@ist.psu.edu

Cyberinfrastructure or e-science has become crucial for scientific progress and open source systems have greatly facilitated design and implementation. In chemistry, the growth of data has been explosive and timely and effective information and data access is critical. We discuss our ChemXSeer (funded by NSF Chemistry) architecture, a portal and search engine for academic researchers in environmental chemistry, which integrates the scientific literature with experimental, analytical and simulation datasets. ChemXSeer consists of information crawled from the web, manual submission of scientific documents and user submitted datasets, as well as scientific documents and metadata provided by major publishers. Information gathered from the web is publicly accessible whereas access to restricted publisher resources will be provided by linking to their respective sites and users can control access to their data. Thus, instead of being a fully open search engine and repository, ChemXSeer will be a hybrid one, limiting access to some resources.

ChemXSeer offers some unique aspects of search not yet present in other scientific search services or search engines. We have developed or are developing algorithms for the extraction of tables, figures, and chemical names and formulae from scientific documents enabling users to search on those fields. In particular ChemXSeer will provide the following search features:

- Full text search
- Author, affiliation, title and venue search
- Table search
- Figure search
- Chemical formulae and name search
- Citation and acknowledgement search
- Citation linking and statistics

ChemXSeer takes advantage of many open source search and indexing tools such as Lucene and CiteSeer. For dataset search, we are developing tools that automatically annotate published data representations such as figures that permit researchers to annotate their datasets by providing both document-level and attribute-level metadata in OAI-PMH format. This level of data annotation permits more effective data search both at the attribute and semantic levels, and allows browsing of datasets and linking to existing scientific literature and other datasets in our and other repositories.

Because ChemXSeer requires unique information extraction, several different machine learning methods, such as conditional random fields, support vector machines, mutual information based feature selection, sequence mining, are critical for performance. We give a progress report on ChemXSeer and draw lessons for other e-science and cyberinfrastructure systems in terms of design, implementation and research.

Topic: Machine Learning Applications, Chemistry
Preference: Oral or Poster

References:

[ChemXSeer] http://chemxseer.ist.psu.edu

[Lucene] http://lucene.apache.org

[PubChem] http://pubchem.ncbi.nlm.nih.gov/

[Afeefy 2005] H.Y. Afeefy, J.F. Liebman, and S.E. Stein, "Neutral Thermochemical Data" in NIST Chemistry WebBook, NIST Standard Reference Database Number 69, Eds. P.J. Linstrom and W.G. Mallard, National Institute of Standards and Technology, Gaithersburg MD, 20899, June 2005 (http://webbook.nist.gov).

[Atkins 2003] Daniel E. Atkins, Kelvin K. Droegemeier, Stuart I. Feldman, Hector Garcia-Molina, Michael L. Klein, David G. Messerschmitt, Paul Messina, Jeremiah P. Ostriker, Margaret H. Wright, "Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure," 2003.

[Bolelli 2007] Bolelli, L., Lu, X., Liu, Y., Jaiswal, A., Bai, K., Councill, I., Mitra, P., Wang, J.Z., Mueller, K., Kubicki, J., Garrison, B., Bandstra J., Giles, C.L. "ChemXSeer: A Chemistry Web Portal for Scientific Literature and Datasets" Open Repositories Conference, San Antonio, Texas, 2007.

[Fletcher 1996] Fletcher, D.A., McMeeking, R.F., Parkin, D., J. "The United Kingdom Chemical Database Service", Chem. Inf. Comput. Sci., 36, 746-749, 1996.

[Hey 2006] Tony Hey, Anne E. Trefethen," Cyberinfrastructure for e-Science," Science, 308 (5723): 817-821, 2006.

[Hughes 2004] Gareth Hughes, Hugo Mills, David De Roure, Jeremy G. Frey, Luc Moreau, m. c. schraefel, Graham Smith and Ed Zaluska, "The semantic smart laboratory: a system for supporting the chemical eScientist," Org. Biomol. Chem., 2, 3284 - 3293, 2004.

[Jaiswal 2006] A. Jaiswal, C. L. Giles, P. Mitra, J.Z. Wang, "An Architecture for Creating Collaborative Semantically Capable Scientific Data Sharing Infrastructures," 8th International Workshop on Web Information and Data Management (WIDM2006), Arlington, VA, 2006.

[Liu 2006] Ying Liu, Prasenjit Mitra, C. Lee Giles, Kun Bai, "Automatic extraction of table metadata from digital documents," ACM/IEEE Joint Conference on Digital Libraries 2006 (JCDL 2006): 339-340, 2006.

[Liu 2007] Ying Liu, Kun Bai, Prasenjit Mitra, C. Lee Giles, "TableSeer: automatic table metadata extraction and searching in digital libraries," ACM/IEEE Joint Conference on Digital Libraries (JCDL 2007): 91-100, 2007.

[Lu 2006] Xiaonan Lu, Prasenjit Mitra, James Ze Wang, C. Lee Giles, "Automatic categorization of figures in scientific documents," Joint Conference on Digital Libraries 2006 (JCDL 2006): 129-138, 2006.

[Lu 2007] Xiaonan Lu, James Z. Wang, Prasenjit Mitra and C. Lee Giles, "Automatic Extraction of Data from 2-D Plots in Documents," International Conference on Document Analysis and Recognition (ICDAR 2007), 2007.

[Murray-Rust, 2001] P. Murray-Rust, H. S. Rzepa and M. Wright, "Development of Chemical Markup Language (CML) as a System for Handling Complex Chemical Content," New J. Chem., 618-634, 2001.

[Snow 2006] D.R. Snow, M. Gahegan, C.L. Giles, K.G. Hirth, G.R. Milner, P. Mitra, J.Z. Wang, "Cybertools and Archaeology," Science, 311: 958-959, 2006.

[Sun 2007] B. Sun, Q. Tan, P. Mitra, C.L. Giles, "Extraction and Search of Chemical Formulae in Text Documents on the Web," Proceedings of the 16th International World Wide Web Conference (WWW 2007), 251-260, 2007. (Nominated for best student paper).