Chem$_X$Seer: An eChemistry Web Search Engine and Repository

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[2], Qingzhao Tan[2], Anuj Jaiswal[1], James Kubicki[4], Barbara Garrison[3], Joel Bandstra[3]
[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

Abstract

With the amount of scientific digital content available online constantly increasing, the community of science has been increasing its efforts towards automatically collecting and organizing such information. This has led to such system architectures as CiteSeer , a digital library and search engine focusing on scientific literature in Computer and Information Science and related fields, the virtual observatory for astronomy, arXiv, and others. In chemistry, the growth of data and collaboratory teams has been explosive. This has led to the Chem$_X$Seer project and architecture, a portal and search tool for academic researchers in environmental chemistry that integrates the scientific literature with experimental, analytical and simulation result datasets.

1. Introduction

E-science or cyberinfrastructure have 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 [Atkins 2003, Hey 2006]. Many have argued that cyberinfrastructures for science are domain sensitive [Snow 2006] and many have been proposed. We have proposed and are developing the Chem$_X$Seer architecture, a portal for academic researchers in environmental chemistry, which integrates the scientific literature with experimental, analytical and simulation datasets. Chem$_X$Seer will be comprised 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 crawled by Chem$_X$Seer from the web and user submitted data will be publicly accessible whereas access to publisher resources can be provided by linking to their respective sites. Thus, instead of being a fully open search engine and repository, Chem$_X$Seer will be a hybrid one, limiting access to some resources.

2. Chemistry and Cyberinfrastructure

Chemical research is becoming increasingly collaborative in scope and approach. For example, within the Penn State Center for Environmental Kinetics Analysis (CEKA), researchers are taking a multi-disciplinary approach to linking kinetic information in environmental chemistry across spatial and temporal scales (see www.ceka.psu.edu for more information). A main goal of such research is to integrate experimental, analytical, and simulation results performed on systems from molecular to field scales in order to better approximate the complex physical, chemical, and biological interactions controlling the fate and transport of contaminants. Researchers in this area have realized for some time that new scientific questions can be generated when users have access to a broad spectrum of related results. Subsequently, as connections are made among field observations, experimental kinetic results, spectroscopic analyses, and model predictions, gaps in the information web will become apparent. Approaches to filling these gaps can then be addressed by the collaborative team. In addition, when inconsistencies among results are identified through use of databases and collaboration tools, determining the reasons for these differences can motivate and guide researchers to formulate experiments to eliminate discrepancies or make new discoveries. Consequently, CEKA represents an ideal base and test bed in the development of databases and associated collaboratory tools that will improve communication among scientists working in various disciplines and at vastly different scales. An easily queried, intelligent database would provide access to critically relevant data for a diverse community of users, enabling these users to achieve higher order scientific

goals. In short, data collection and synthesis will lead to better science and improved education of scientists.

Academic chemistry is now focusing on cyberinfrastructure in many areas, with a focus on building tools that are open source and shared within their community. There is much interest on chemistry web services [Coles 2005], chemical markup languages [Murray-Rust 2001], chemical ontologies [Chebi], chemical literature indexing and search [PubChem], data repositories [Fletcher 1996, Afeefy 2005], echemistry notebooks [Hughes 2004], and others. As part of this wave of development, the Chem$_X$Seer system will focus on the needs of domain scientists in chemistry, especially the interdisciplinary needs that arise when addressing complex kinetics synthesis problems.

3. The Chem$_X$Seer System

Chem$_X$Seer intends to offer unique aspects of search not yet present in other scientific search services. We have developed or are developing algorithms for the extraction of tables, figures, equations and formulae from scientific documents enabling users to search on those fields. Chem$_X$Seer intends to or currently does provide the following search features:

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

A beta version of Chem$_X$Seer is online and working.

4. Data Search

For dataset search, we are developing tools that automatically annotate published data representations such as figures and that permit researchers to annotate their datasets by providing both document-level and attribute-level metadata in OAI-PMH format [Jaiswal 2006]. This level of data annotation provides us with the opportunity of searching data more effectively both at the attribute and semantic levels, browsing datasets and linking to existing scientific literature and other datasets in our and other repositories. The data search feature is now available as a beta in Chem$_X$Seer

5. Chemical Formula Search

Current search engines do not recognize chemical entities (chemical names and formulae). A scientist, who seeks to search for information related to chemical molecules from text documents cannot do so in any meaningful way except performing exact keyword search. We show how a chemical-entity-aware search engine can be designed and built and demonstrate empirically that it improves the relevance of the returned results for search queries involving chemical entities. Our search engine first extracts chemical entities from text, performs a novel indexing suitable for chemical formulae and names, and supports different query models that a scientist may require. We develop a model of hierarchical conditional random fields (HCRFs) for entity tagging that considers long-term dependencies at the higher (hierarchical) levels. We create an algorithm of independent frequent subsequence mining (IFSM) to discover sub-terms of chemical names and estimate their probabilities of occurrence. We also develop an unsupervised hierarchical text segmentation (HTS) method to represent a sequence with a tree which is based on discovered independent frequent subsequences. Finally, we create unique formulae ranking algorithms for formulae search. We introduce various query models for chemical name searches. Experiments and examples show that our approach performs reasonably well [Sun 2007]. A beta of formula search is now available in Chem$_X$Seer.

6. Table Search

Tables are ubiquitous in scientific documents where they are widely used to present experimental results or statistical data in a condensed fashion. However, current search engines do not support table search. Scientists and scholars often extract data from tables in documents by hand. We have created a table search engine (TableSeer) that eliminates that burden enabling users to automatically search for and examine tables. TableSeer crawls scientific documents, detects tables from documents, extracts and indexes tables metadata, and enables end-users to search for tables. We also propose an extensive set of medium-independent metadata for tables because there exists no tablemarkup language and universal metadata specification for tables that scientists and other users can adopt for representing table information in documents. Given a query, TableSeer ranks the matched tables using an innovative ranking algorithm, TableRank, which also calculates the relevance between a table and a query by combining the query-dependent features of the table with the query-independent features of the document where the table exists. Overall, the ranking scores are calculated using a specific term-weighting scheme that aggregates multiple impact factors from three levels: the term level, the table level, and the document level. We demonstrate the value of our specialized table-search engine by empirical studies using scientific documents [Liu 2006, 2007]. A beta of TableSeer is now available in Chem$_X$Seer.

7. Figure Search

Like tables, much important information in documents resides in figures. We are interested in all of figure search. However, our current focus is on searching for and extracting data from plots. Two-dimensional (2-D) plots in digital documents contain important information. Often, the results of scientific experiments and performance of businesses are summarized using plots. Although 2-D plots are easily understood by users, current search engines rarely utilize the information contained in the plots to enhance the results returned in response to queries. We have developed an automated algorithm for extracting information from line curves in 2-D plots. The extracted information can be stored in a database and indexed to answer end-user queries and enhance search results. We have collected 2-D plot images from a variety of resources and tested our extraction algorithms. Experimental evaluation has demonstrated that our method can produce results suitable for real world use [Lu 2006, Lu 2007]. We intend to make a beta for plot search available in Chem$_X$Seer.

7. Search Integration

Integration of search results is a nontrivial issue. We propose to investigate new methods for integrating chemical search based on molecular centric views permitting users to have easy access to all related information [Bolelli 2007]. We also intend to link search results to other available databases.

8. Conclusions

Chem$_X$Seer is an ongoing cyberinfrastructure project in chemistry. By offering chemists automatic access to information and data not previously available, Chem$_X$Seer will change the way chemists do research and will significantly enhance scientific productivity and contribute to new discoveries.

9. Acknowledgements

The Chem$_X$Seer cyberinfrastructure research and project is supported in part by a grant from NSF Chemistry. We give special thanks to the programming support of Juan Pablo Fernandez Ramirez.

10. References

[Chebi] http://www.ebi.ac.uk/chebi/

[Chem$_X$Seer] http://chemxseer.ist.psu.edu

[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. "Chem$_X$Seer: A Chemistry Web Portal for Scientific Literature and Datasets" Open Repositories Conference, San Antonio, Texas, 2007.

[Coles 2005] Simon J. Coles, Nick E. Day, Peter Murray-Rust, Henry S. Rzepa and Yong Zhang, "Enhancement of the chemical semantic web through the use of InChI identifiers," Org. Biomol. Chem., 3, 1832 – 1834, 2005.

[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.

[Murray-Rust, 2004] Peter Murray-Rust, Henry S. Rzepa, S.M. Tyrrell, Yong Zhang, “Representation and use of Chemistry in the Global Electronic Age,” Org. Biomol. Chem., 10:1039, 2004.

[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 award.)