

GEONSearch: From Searching to Recommending

Ullas Nambiar, Bertram Ludaescher

Department of Computer Science,
University of California, Davis

Ghulam Memon, Dogan Seber

San Diego Supercomputing Center,
University of California, San Diego

Abstract

The Geosciences Network (GEON) project is a large-scale collaborative effort aimed at creating cyberinfrastructure for the Earth Sciences. GEON is a repository of heterogeneous data which can be integrated, queried, analyzed and visualized by geo-scientists with GEON-provided ontology enabled tools. A critical first step in efficient retrieval and integration of data from GEON is the ability to identify the datasets of interest. GEONSearch - a component of the GEON Portal, plays a very important role in GEON by enabling users to discover datasets of interest by searching over *Title* and *Description* of the dataset, *spatial* and *temporal* metadata and mapped concepts from GEON ontology. While GEONSearch does enable searching over many associated features, the underlying search model is keyword based. Specifically, GEONSearch currently returns only datasets whose features explicitly match the keywords provided by the user. Thus effective retrieval of datasets from GEONSearch will require users to construct keyword queries that clearly identifies their need - a task that is often difficult even for experts. By focussing on exact keyword matches GEONSearch would miss to provide datasets that are relevant to the user query but whose metadata do not match with they query. For example, if users search for *California*, they will not find relevant documents mentioning only *San Diego*. Similar problems could arise when searching over *temporal* metadata. Concept level search provided by GEONSearch could help overcome this problem to a certain degree. However, effective use of concept-based search will require user to formulate the query in terms of available concepts. Hence, our focus is in improving the relevance of results provided by the easy-to-use keyword search currently provided by GEONSearch. We will do so by incorporating background knowledge available from GEON ontologies, dataset-to-ontology mappings, past usage information and content (schema and data) overlap between sources. The primary motivation behind this talk is to avoid the *frog in a well* syndrome of doing research where the tools developed do not meet user needs. We provide an overview of the research being done to improve GEONSearch with the intend of receiving feedback from the users - geoscientists. In particular, we are interested in obtaining *use-cases* that will be helpful to our research and/or bring out additional challenges to be solved.

In this talk, we will present results of ongoing research to allow GEONSearch to learn relevance of a dataset to a query, to other datasets and to GEON ontologies. With this knowledge GEONSearch will be able to recommend related datasets and ontologies to users once they pick a dataset of interest. More importantly, GEONSearch would be able to provide justifications for its recommendations that are based on user accepted semantics. This ability to learn related/relevant datasets would also be useful in improving the data registration process by allowing GEON to suggest and/or automatically map datasets to relevant ontologies which the user may have missed.