Abstract

The rapid growth of scientific data shows no sign of abating. This growth has led to a new problem: with so much scientific data at hand, stored in thousands of datasets, how can scientists find the datasets most relevant to their research interests? We have addressed this problem by adapting information retrieval techniques, developed for searching text documents, into the world of (primarily numeric) scientific data. We propose an approach that uses a blend of automated and “semi-curated” methods to extract metadata from large archives of scientific data, then evaluates ranked searches over this metadata. We describe a challenge identified during an implementation of our approach: the large and expanding list of environmental variables captured by the archive do not match the list of environmental variables in the minds of the scientists. We briefly characterize the problem and describe our initial thoughts on resolving it.

Prior Work

Addressed the problem of finding relevant data in a “big data” archive (Megler and Maier, 2011)
- Many datasets, dataset shapes and sizes, physical locations, formats, tools
- “Misremembered” datasets → lost data
- Example information need:
  - “observations collected near [lat = 45.5, lon = -124.4] in mid-2010, with temperature between 5-10°C”

Solution: Build search engine for scientific data

IR Architecture Adapted to Scientific Data Search

Ranked Search Over Data: Location, Time, Variables

• Build metadata catalog to represent archive contents
• Individual datasets scanned once, summarized into a “feature” per data
• Features stored in catalog
• Similarity search is performed over catalog contents
• Search results ranked on distance based similarity to query terms

Motivation

Emerging problem: Many names for same environmental variable

“Semantic diversity”

“Similar problems in other areas, e.g. units

The Metadata Wrangling Process

- Scan archive
- Add external metadata
- Generate hierarchies
- Perform known transformations

Example Process

• Set of composable components
• Compose into “metadata processing chain”
• Details of process different for each archive

Discovering Transformations with Google Refine

- Extract catalog entries to Google Refine
- Export transformation rules
- Transform against metadata

For More Information

Contact V.M. Megler at: vmegler@cs.pdx.edu or David Maier at: maier@cs.pdx.edu

With thanks to the scientists at Center for Coastal Margin Observation and Prediction (CMOP). This work is supported by NSF award OCE-0424602