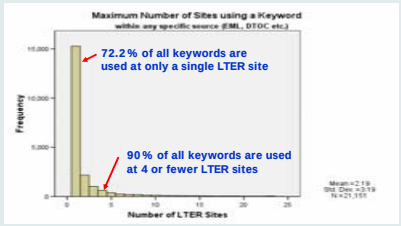
AHIS Dataset Search Engine: An intelligent approach to EML Data Management

Hung V. Nguyen, Corinna Gries, and Hasan Davulcu

Overview

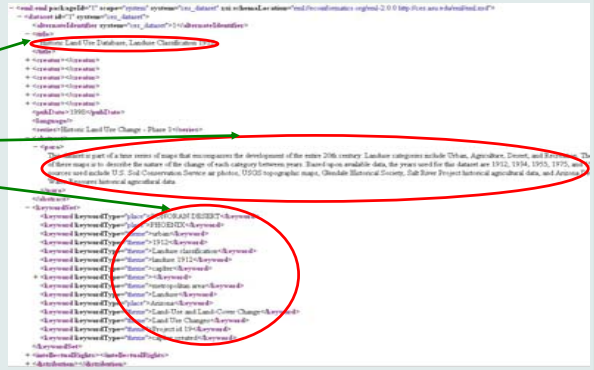
Story:

- Ecological data is widely organized and documented in **E**cological **M**arkup **L**anguage – EML format. Efficient dataset retrieval system needs to understand the important keywords and relations among them. A large collection of terms from **L**ong **T**erm **E**cological **R**esearch (LTER) is collected
- Some times a user's search query may be an imperfect description of their information need. Even when the information need is well described, a search engine or information retrieval system may not be able to retrieve documents matching the query as stated. In this project, we develop a search engine empowered with text mining techniques for ecological datasets to bridge this gap



Technique:

- **Data extraction:** Extract informative keywords from *important parts* of a dataset, namely, abstract, keyword list, title.
  Stop words (e.g. and, or, the, an, etc) are pruned
- **Text Mining:** Study the statistics of terms correlation will help to achieve:
  - Higher accuracy in suggesting related terms for users
  - Higher accuracy in relevance of retrieved datasets



Search Engine

Process the user's queries

Return the relevant datasets

Return the related keyword list to each query term (using indexing engine)

Data Loading Engine

Setup connection to Exist Database

Update the searching repository with new datasets

Extract important keywords from new datasets

Filtering Tool

Allow users to prune irrelevant or meaningless keywords

Update the searching repository with updated keyword lists

Indexing Engine

Construct Term-Document (term-doc) matrix M:

$M_{i,j}$ = frequency of term *i* in document *j*

Compute the term correlation matrix S

$S = M * M^T$

Normalize S to obtain C

$$C_{u,v} = \frac{S_{u,v}}{S_{u,u} + S_{v,v} - S_{u,v}}$$

Overview Architecture and Work Flow of the System

Screenshot Example

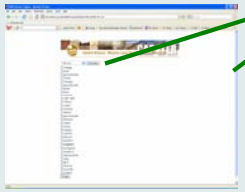
"water"

Our system offers a flexibility for users to either narrow down or expand the search for relevant datasets by building boolean search based on "contextual-aware" related terms of current query term(s).

"water **AND** maricopa"

The human evaluation shows that our approach yields high accuracy of relevance in terms of both related keywords and datasets.

"(water **AND** maricopa) **OR** ecosystem"



Our system also offers a set of Web-Based applications that let the users update new data from EML Exist database, prune irrelevant keywords and re-index filtered keyword set for search engine

Contact:

Hung V. Nguyen is a Ph.D. candidate at CSE Department, ASU.
His research interests are Web/Text Mining, Web Advertising, Machine Learning and Applications in Data Mining.
Email: hung@asu.edu
Web: http://www.public.asu.edu/~vnguye1

Gries received her Ph.D in 1988. Currently, she is information manager at the Central Arizona - Phoenix LTER site and co-chair of the LTER information managers committee. Her research interests are the implementation of community standards for discovery, access, visualization and use of ecological data, development of large natural history collections databases and online collection management tools. She is leading development and implementation of Arizona Hydrologic Information System as a pilot project on the State level and a node to CUAHSI and the USGS Geologic Information System.
Email: corinna@asu.edu

Hasan Davulcu is an Asst. Professor at CSE Department, ASU.
Dr. Davulcu directs CIPS Lab. Lab's research focuses on developing novel data mining techniques and tools for structuring and organizing unstructured sources such as text, Web and biological data into semantic machine processable information. His interests also include Workflows, Web Services and Databases Systems.
Email: hdavulcu@asu.edu
Web: http://www.public.asu.edu/~hdavulcu