Advanced Search Technologies for Unfamiliar Metadata

Fredric Gey
UC Data Archive and Technical Assistance, MS 5100
Hui-Min Chen
Barbara Norgard
Michael Buckland
Youngin Kim
Aitao Chen
Byron Lam
Jacek Purat
Ray Larson
School of Information Management and Systems, MS 4600
University of California, Berkeley
Berkeley, California, 94720
(gey@ucdata.berkeley.edu;,
(hmchen,barbara,buckland,kimy,aitao,byronlam,purat,ray)@sims.berkeley.edu)

ABSTRACT

Searching of databases (textual or numeric) is likely to be effective and efficient only if the user is familiar with the classification, categorizing, and indexing schemes (metadata vocabularies) being searched. Therefore, it is obviously beneficial to provide a bridge between the user's ordinary language and the metadata vocabularies of the unfamiliar database in order to compensate for abbreviated, cryptic, or specialized terminologies. Advanced search technologies would utilize customized "Entry Vocabulary Modules" (EVM) which respond adaptively to the user's ordinary language query with a ranked list of search terms in the target metadata vocabularies that may more accurately represent what is sought in the unfamiliar database. These "Entry Vocabulary Modules" can serve both as an indexing device and a search aid. This project has developed EVMs for several metadata vocabularies, including domestic and international patent classifications and U.S. Standard Industrial Classification Codes. An agent-based architecture is under design to lighten the task of cracking alien metadata vocabularies.

1.0 INTRODUCTION

A wealth of databases, whose content is textual, numeric and mixed are now appearing on the internet through the World Wide Web. Indeed, classical bibliographic search companies such as DIALOG have now developed web interfaces (http://www.dialogweb.com). Search technology to locate and retrieve these databases is currently quite primitive, consisting primarily of internet search engines or local webcrawlers to find potentially relevant web pages which serve as entry points to complex (and heterogeneous) database applications. However, each of these database

applications has its own idiosyncratic metadata which describes the structure and detailed content of the database, and whose description may not (indeed, usually will not) correspond to the ordinary language search terms submitted by the less experienced searcher.
For example, a user searching for "Automobile" in the Federal Trade Import Export database (http://govinfo.kerr.orst.edu/impexp.html) will receive a 'No Results on Search for "automobile"' message and she is unlikely to find the correct search term abbreviation "PASS MTR VEH, SPARK IGN ENG" which would actually retrieve this import/export category.

An alternative is for the user to be familiar with the technical terminology and abbreviations of a multi-digit numeric hierarchical category system and to navigate down from equally obscure categories. The hierarchy for the unfamiliar vocabulary term "computers" in the International Harmonized Commodity Classification System (which classifies U.S. Imports and Exports) proceeds from

- *HS 84 (Nuclear reactors, boilers, machines and mechanical appliances)* to
- *HS 8471 (Automatic data processing machines and units thereof, magnetic or optical readers, machines for transcribing data)* to
- *HS 84712 ("Digital Auto data proc mach contng in the same housing a CPU and input&output device")*

Our DARPA-funded research project has been developing advanced search capabilities to discover and navigate unfamiliar metadata. Using what is known as "Entry Vocabulary Modules" we create associations between ordinary language and domain-specific technical metadata vocabulary used to describe databases. During the first year the project has developed several prototype entry vocabularies to map from ordinary language to Library of Congress Classification, ordinary language to U.S. Patent Classification, ordinary language to BIOSIS (biological abstracts) concept codes and ordinary language and to INSPEC thesaurus terms.
(http://www.sims.berkeley.edu/research/metadata/oasis.html).

Currently the project is developing an entry vocabulary module to map from ordinary language to Foreign Trade Import/Export data classified by the International Harmonized Commodity Classification System (16,000 categories of commodities).

2.0 METHODOLOGY

The process of creating an entry vocabulary module is one of Bayesian inference, wherein sufficient training data (consisting of document texts) are downloaded (usually in MARC record formats using the Z39.50 protocol for efficiency) from a document database to provide a probabilistic matching between ordinary language terms and the specific metadata classifications which have been used to organize the (either textual or numeric) data. Developing the entry vocabulary utilizes both natural language processing modules [Kim & Norgard 1998] as well as statistical language techniques for extracting key phrases (e.g. 'ink jet printer') to map to specialized classifications.

The particular technique of creating a ranked list of probably relevant terms in the target metadata vocabulary from any given searcher input was developed under the name Classification clustering" by Ray Larson [Larson 1991]. The method utilized a probabilistic interpretation of vector-spaced retrieval, coupled with a logistic regression ranking method [Gey 1994]. A two-stage lexical collocation process is used. The first stage is *creation* of an Entry Vocabulary Module, a "dictionary" of associations between the lexical items found in the titles, authors, and/or abstracts and the metadata vocabulary (i.e. the category codes, classification numbers, or thesaural terms assigned), using a likelihood ratio statistic as a measure of association. In the second stage, *deployment*, the dictionary is used to predict which of the metadata terms best represent the topic represented by the searcher's terms [Plaunt & Norgard, 1998].

We designate the natural language vocabulary to consist of "A" terms (words or phrases used by the searchers) which are to be mapped to the "B" terms of the specialized metadata vocabulary. To calculate the degree of association between A-terms and B-terms, we utilize a contingency table which relates documents in which A-terms are found with B-terms which have been used to index the documents. In information retrieval literature, contingency tables are commonly used as a way to measure the relative frequency of an event involving two entities . Abstractly frequencies for four combinations of a A-term and a B-term are calculated and placed the cell of the contingency table as shown below.

Table 1. The contingency table for a pair of A-term and B-term

$\lvert A \cap B \rvert$	$\lvert A \cap \bar{B} \rvert$	|A|
$\lvert \bar{A} \cap B \rvert$	$\lvert \bar{A} \cap \bar{B} \rvert$	$\lvert \bar{A} \rvert$
|B|	$\lvert \bar{B} \rvert$	N

$\lvert A \cap B \rvert$ is the frequency of the A-term and the B-term pair, $\lvert A \cap \bar{B} \rvert$ the frequency of the A-term but not the B-term pair, $\lvert \bar{A} \cap B \rvert$ the frequency of the B-term but not the A-term pair, and $\lvert \bar{A} \cap \bar{B} \rvert$ the frequency of neither the A-term nor the B-term pair. In addition, |A| is the frequency counts for the A-term, $\lvert \bar{A} \rvert$ the frequency counts for pairs without the A-term, |B| the frequency counts for the B-term, $\lvert \bar{B} \rvert$ the frequency counts for pairs without B-term, and N the frequency counts for each unique pair.

Whenever an association is to be measured, standard statistical tests such as χ^2 test and Z-score tests are usually taken for granted. However, one of the problems in applying some standard statistical tests to the text is the assumption that the words of the text conform to these distributions. It has long been recognized, in the information retrieval literature, that the frequency of words in texts follows a Zipfian distribution: Rank * Frequency $\approx$ Constant [Luhn, 1958; Salton, 1989]. This means that the so-called rare events (on the tails of a distribution curve) in texts are very common. Worse, the importance of such “rare” occurrences are greatly overestimated by tests which assume a normal distribution.

A binomial event counting model where each event A being counted is compared to a particular event B, is more appropriate for the association dictionary building process. It provides a chance to measure the independence of event A and B, and further calculate the degree of association between event A and B. For this purpose, the likelihood ratio for comparing these two binomial events as developed by Dunning [Dunning 1993] is adopted and incorporated in the design of builder agents. These associations are ranked by a log frequency weight. See [Plaunt and Norgard 1998] for details.

4.0 SEARCHING NUMERIC CLASSICATIONS -- U.S. Standard Industrial Classification

Standard Industrial Classification (SIC) codes have been used by the United States Government for about 50 years. The SIC system is authorized and managed by the Statistical Policy office of the Office of Management and Budget. . The purpose of using SIC codes is to facilitate the comparability of statistical data that describes establishments in the United States economy. The primary intended use of SIC codes to collect, tabulate, and publish establishment data by industry. SIC codes have the following uses

- To *aggregate data*: SIC codes are used to facilitate use of data by aggregating otherwise unmanageable detail. This is one of their most important functions.
- To *aid sampling*: SIC codes are used in sampling as a way to collapse commodity detail into aggregated estimates.
- To *facilitate data comparison*: SIC codes are used to make comparisons between data sets.

4.1 The Structure of SIC

The SIC system is hierarchical. Each level of the system provides an aggregation of detail found at the next lower level. The top 11 divisions are divided into 83 2-digit major industry groups. These are further subdivided into 416 3-digit industry groups, which break down into 1,005 4-digit industries. Most data available to the public (such as the County Business Patterns) is recorded at the 4-digit level. More detailed product information is recorded with 5-digit product class codes and 7-digit product codes.

The top level of the SIC consists of the following categories:

Division A. Agriculture, forestry, and fishing
Division B. Mining
Division C. Construction
Division D. Manufacturing
Division E. Transportation, communication, electric, gas,sanitary services
Division F. Wholesale trade
Division G. Retail trade
Division H. Finance, insurance, and real estate
Division I. Services
Division J. Public administration
Division K. Nonclassifiable establishments

4.2 Example SICs and their search

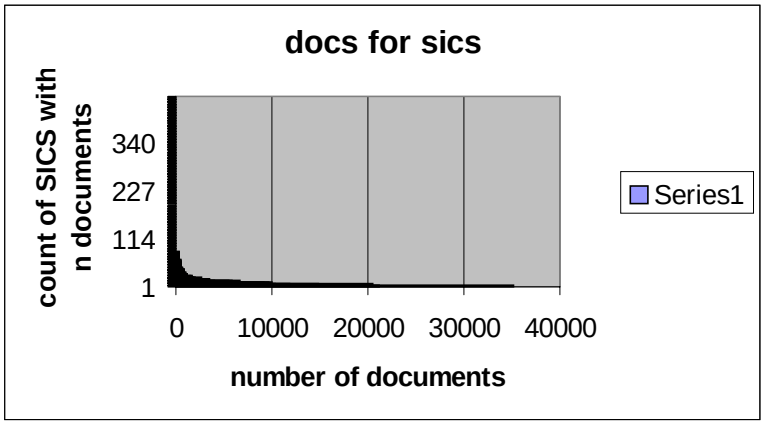
An online version of SIC searching can be found at (http://www.lib.virginia.edu/socsci/sic.html) . Yet, if one does a search for the phrase 'ink jet printers' in this engine, we obtain no classifications, even though ink jet printers for personal computers have become a major product category in the past decade. In preparation for mapping between foreign trade data classifications and natural language, we identified those SIC classifications associated with foreign trade categories. Of the 16,000 HS categories we identified 448 SICs which were associated with them. A database of trade magazines was identified for which many articles have been hand classified by SIC. From this database we downloaded a 246,423 document dataset for training purposes and created an entry vocabulary for the SIC (http://metaphor.sims.berkeley.edu/oasis/sic.html). If one does the search 'ink jet printers' for this vocabulary, one obtains the following ranked classification

- 3579 OFFICE MACHINES, NSPF, AND PARTS, NSPF
- 3825 INSTRUMENTS FOR MEASURING AND TESTING ELECTRICITY AND ELECT SIGNALS
- 2893 PRINTING INKS

and the searcher is free to choose that category which best reflects his information need.

4.3 Skewedness of the training distribution

One of the problems encountered in developing the SIC entry vocabulary module was a skewedness of the distribution of training documents associated with particular SICs. For example in the training database described above, 130 classifications, or nearly 1/3 of the entire set, had fewer than 20 documents to train on. At the other end, the most frequently occurring SIC, 3571 ELECTRONIC COMPUTERS had 35,230 documents, and the top 20 SICs had 187,557 documents or 76 percent of the collection.(see graph)



The initial version of the EVM yielded SIC 3571 for many inquiries, including 'automobile' because of the prevalence of article about the use of computer controlled engine and transmission environments in automobiles. Our final EVM for the SIC normalized the distribution to achieve approximately equal numbers of training documents randomly selected for each SIC. The optimal solution to this real-life problem of skewed training distributions remains an open research problem.

5.0 AGENT ARCHITECTURE FOR METADATA VOCABULARY

During the past year the project has devised an agent architecture (Figure1) consisting of multiple interacting agents which "divide and conquer" the entry vocabulary technology tasks. These range from directory recognition and search agents which locate and identify databases through builder agents which create associations from training databases downloaded by data retrieval agents. Desktop agents help the user to define domains of interest and deploy the association dictionaries created by entry vocabulary agents.

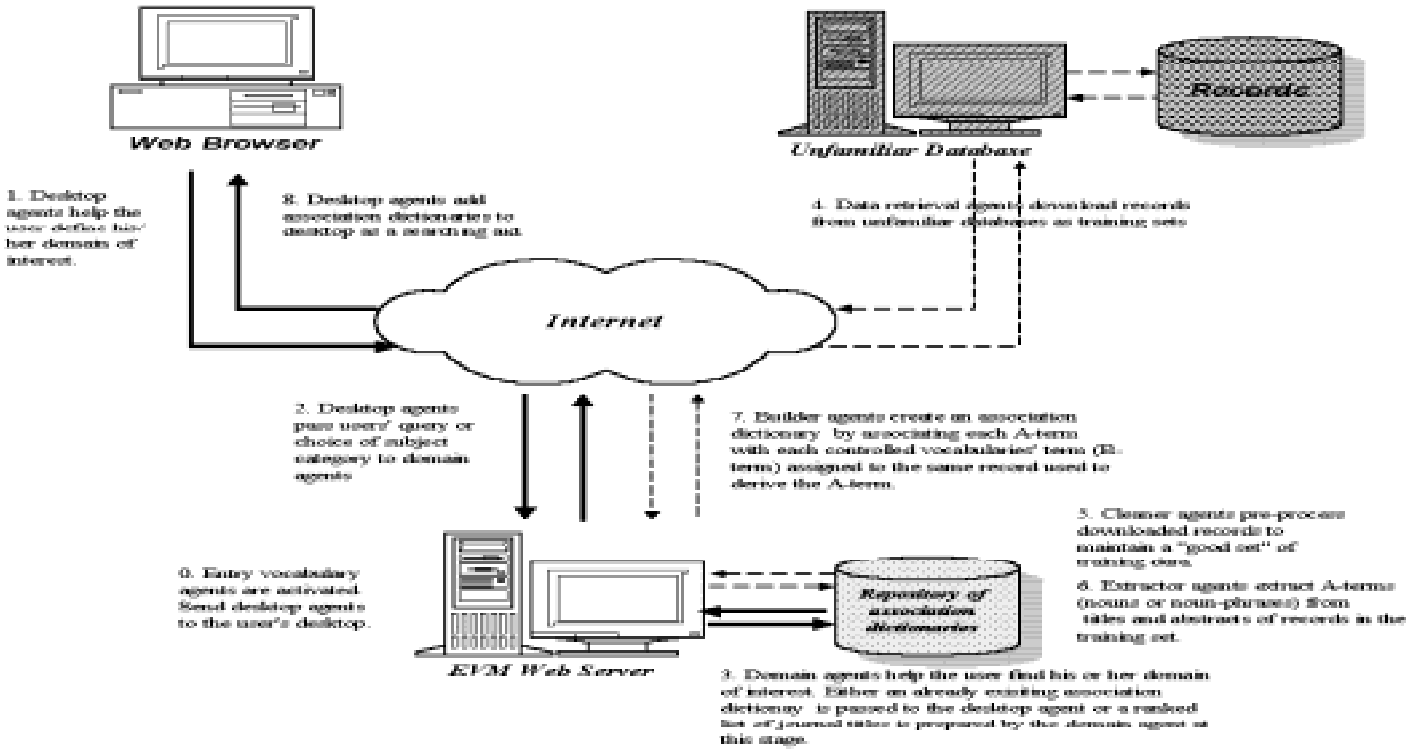


Figure 1. Architecture of Multi-Agent-Based Search Support for Unfamiliar Databases

Running on a UNIX platform, there are seven kinds of agents in the EVM system, entry vocabulary agents, desktop agents, domain agents, data retrieval agents, cleaner agents, extractor agents, and builder agents. Agents are designed to possess a different scope of mission, authority, autonomy, resources, and life expectancy. The EVM agents are introduced below in the order in which they come into play during a search session.

5.1 Entry vocabulary agents

Entry vocabulary agents are at the heart of the EVM system. They are always alive during a session starting from the initial acceptance of a user's request and ending with a session termination signal passed by desktop agents. To request a search service, the user enters the URL of the server computer where the EVM system resides such as

http://www.sims.berkeley.edu/research/metadata/index.html When the server accepts a request, it initiates a process in which a clone of entry vocabulary agents is made and activated. The activated entry vocabulary agent functions as the master agent of the search session.

5.2 Desktop agents

The desktop agent provides a graphical user interface that links the user with the entry vocabulary agent. It is the only type of agent that resides on the client computer to communicate with the user directly. In its role as assistant to the user, the desktop agent stays alive during the whole search session. The major tasks of the desktop agent are to:

1. Gather the user's instructions and search queries.
2. Examine the validity of the user's input.
3. Pass the user's instructions to the entry vocabulary agent.
4. Forward the user's input as well as search queries to other agents such as the domain agent.
5. Acquire the status of other agents.
6. Display and update the status of other agents in a real-time fashion.
7. Display association dictionary building progress on the screen in real time.
8. Add the resulting association dictionaries on the desktop to aid the user searching the database.
9. Provide online help to guide the user specifying his or her domain of interest.
10. Help the user browsing the resulting association dictionaries.
11. Detect the termination of a search session.

5.3 Domain agents

The task of domain agents is to prepare ranked list of journal titles that reflect the user's domain of interest. Ordinarily, metadata vocabularies are studied as a whole and the entry vocabulary modules can be designed for a whole database with its own metadata vocabulary. However, in practice, users are rarely equally interested in all of the contents of the whole database. Instead, they are usually interested in some specific domain reflecting their particular interest in a search. Thus, it will be more efficient and cost-effective to concentrate on Entry Vocabulary Modules of topical, work-related domains.

5.4 Data retrieval agents

Data retrieval agents are designed to download records from the unfamiliar database based on the ranked list of journal titles prepared by domain agents. Those records then serve as the training set to create an association dictionary. Currently, records used to build association dictionaries are typically acquired by downloading sets of MELVYL records retrieved from a query specifying a topic of discussion in a domain (An exception to this approach has been the U.S. patents records that require the application of another set of procedures.). MELVYL (http://www.melvyl.ucop.edu/) is an online catalog system of the University of California. It

provides a uniform interface to a variety of databases such as INSPEC, BIOSIS, MEDLINE PLUS, and so forth.

5.5 Cleaner agents

The multiplicity of formatting, even SGML formatting of text, implies a difficult and demanding job of standardizing formats to be understood and prepared for processing by other agents. The records in the downloaded set must transformed or filtered before the actual entry vocabulary can be constructed. The cleaner agent, as its name implies, takes care of this nasty job. It is an open question as to whether each new database being accessed will require some customized code for processing.

5.6 Extractor agents

For each record in the training set, terms (nouns or noun-phrases) from the tagged title and abstract are extracted by the extractor agent. They are A-terms. Then each A-term is paired with the controlled vocabularies' terms, called B-terms, assigned to that record. The extractor agent ends with a list of A-term and B-term pairs for each record in the training set

5.7 Builder agents

Now, it is time for the builder agent to come into play. The major task of the builder agent is to compute the association between the A-term and the B-term in a pair in order to complete the steps for creating a dictionary of ordinary language terms associated with controlled vocabularies' terms. It makes no difference to the builder agent in building a word-based dictionary or phrase-based dictionary because they have exactly the same procedure.

5.8 Implementing the agent architecture

Implementing an agent architecture for as complex a series of tasks as search engine technology for metadata vocabularies is a challenging endeavor, not only for the specification of agent functionality, but also for the specification of inter-agent communication. Our current partial implementations consist of PERL cgi programs with communication via file formats rather than protocols. We are currently investigating appropriate representations and languages for knowledge communication between our EVM agent components.

6.0 CONCLUSIONS AND FUTURE WORK

This paper has presented a novel method for searching unfamiliar metadata vocabularies. Instead of relying on incomplete textual descriptions of the specialized vocabulary being searched, we provide *Entry Vocabulary Modules (EVMs)* to bridge the gap between ordinary searcher and specialized languages. The technique has been applied not only to textual databases found in the literature, but also to numeric databases organized by complex, hierarchical classification schemes. The latter has been demonstrated with a U.S. Standard Industrial Classification module, and will be extended in the future to the

International Harmonized Commodity Classification Scheme (HS), which is used to report Import-Export activities between the United States and countries of the world. The project has designed an *agent architecture* which breaks down the tasks of developing and deploying EVMs into manageable independent components.

Implementation of EVMs has uncovered some fundamental problems in text categorization relating to the skewedness of the training sets in real-life applications. We are researching new algorithms for mitigation of such non-uniform distributions to achieve greater accuracy in the classification and mapping task. We have also investigated whether the technique also has application to cross-language retrieval, where metadata classifications in one language can be mapped to documents in another language which have been indexed using the original language's metadata.

This work is supported by DARPA contract DARPA Contract N66001-97-C-8541;AO# F477: Search Support for Unfamiliar Metadata Vocabularies.

7.0 REFERENCES

[Dunning1993] Dunning, Ted 1993 Accurate Methods for the Statistics of Surprise and Coincidence, Computational Linguistics, vol 19, no. 2, pp.61-74.

[Gey 1994] Gey, F. 1994. Inferring Probability of Relevance Using the Method of Logistic Regression. In: *Proceedings of SIGIR94, the 17th annual ACM conference on Research and Development in Information Retrieval*, Dublin, Ireland, July 4-6, 1994, pp. 222-231.

[Kim & Norgard 1998] Kim, Youngin and Norgard, Barbara.1998. Adding Natural Language Processing Techniques to the Entry Vocabulary Module Building Process. Technical Report. http://www.sims.berkeley.edu/research/metadata/nlptech.html

[Larson 1991] Larson, R. R. 1991. Classification Clustering, Probabilistic Information Retrieval and the Online Catalog, *Library Quarterly*, vol. 61, no. 2 (April), 1991, pp. 133-173.

[Luhn 1958] Luhn, H.P. (1958). Automatic creation of literature abstracts. *IBM Journal of Research and Development,* 2, 159-165.

[Plaunt, C. and Norgard, B.A. 1998]. An association based method for automatic indexing with a controlled vocabulary. *Journal of the American Society for Information Science*, 49(10), 888-902.

[Salton, G. 1989] **Automatic text processing**: **The transformation, analysis, and retrieval of information by computer**. Reading, MA:Addison-Wesley.