Entry Vocabulary – a Technology to Enhance Digital Search

Fredric Gey, Michael Buckland, Aitao Chen and Ray Larson
University of California
Berkeley, CA 94720
gey@ucdata.berkeley.edu, {aitao,buckland,ray}@sims.berkeley.edu

Michael Buckland, Aitao Chen and Ray Larson
School of Information Management and Systems
University of California
South Hall
Berkeley, CA 94720-4600
{aitao,buckland,ray@sims.berkeley.edu

ABSTRACT

This paper describes a search technology which enables improved search across diverse genres of digital objects – documents, patents, cross-language retrieval, numeric data and images. The technology leverages human indexing of objects in specialized domains to provide increased accessibility to non-expert searchers. Our approach is the reverse-engineer text categorization to supply mappings from ordinary language vocabulary to specialist vocabulary by constructing maximum likelihood mappings between words and phrases and classification schemes. This forms the training data or 'entry vocabulary'; subsequently user queries are matched against the entry vocabulary to expand the search universe. The technology has been applied to search of patent databases, numeric economic statistics, and foreign language document collections.

1. INTRODUCTION

The internet has provided a vast and growing amount of searchable information. In the "deep web" (that part of the internet which is not directly searchable using ordinary search engines) we find information deriving from multiple and quite distinct genres, such as images and numeric statistical data as well as textual information in unfamiliar languages. For example United States Foreign Trade Imports and Exports are available at
http://govinfo.kerr.orst.edu/impexp.html.
Data are classified by commodity being shipped, so one can find out, for example, how many purebred Arabian horses were imported to Louisville, Kentucky from Saudi Arabia in any particular month. A commodity search mechanism is provided at this sites to search commodity descriptions associated with the 8,000 commodity codes. However, the search term 'automobile' retrieves nothing, even though we know billions of U.S. dollars of automobile imports enter the United States each year. In order to retrieve automobile imports with using the string search one needs to know that the description is actually "Pass Mtr Veh" an abbreviation for "Passenger Motor Vehicle" and obtain the data shown in Figure 1.

In another case, suppose a searcher from Germany is interested in articles on economic policy and wishes to search in his/her native language. The search term "Wirtschaftspolitik" will likely retrieve documents in German but not English. We need an automatic way to take familiar search terms and map them to unfamiliar terms or classifications without necessarily even knowing what language they were originally expressed in.

We consider vocabulary to be central to search. The vocabulary used by a searcher may not be the same as contained in a document or as the metadata used to classify the document. In order to provide vocabulary mappings, we need to find resources which we can mine for those mappings. Such resources are available in the form of the world's existing electronic library catalogs. If we undertake to mine these resources and we have a technology which can create statistical mappings between vocabularies, we can create *Entry Vocabulary Indexes (EVI)*. **EVI**s are software modules which enhance search by mapping from the users ordinary language to the (sometimes arcane) metadata of the digital resource.

2. ENTRY VOCABULARY TECHNOLOGY

Entry vocabulary technology to create Entry Vocabulary Indexes rests upon four basic components:

- a sufficiently large training set of documents
- a part of speech tagger to identify noun phrases in documents
- software and algorithms to develop probabilistic mappings between words/phrases and metadata classifications
- software to accept search words/phrases and return classifications

In our system we have utilized the Z39.50 protocol to query textual databases located in electronic libraries and download the MARC records which are the results of such queries. Typically, these records are then processed and converted into an XML representation which can be used for further processing and display. The text representation is then (usually, but not always in developing prototypes) processed using a POS tagger such as the Brill tagger [2] and a list of nouns and noun phrases are extracted from each document

Human Language Technology 2001 San Diego, California USA
.

--- U. S. Imports of Merchandise ---

Year	General Imports Quantity	General Imports Customs Value	Imports for Comsumption Quantity	Imports for Comsumption Customs Value
PASS MTR VEH,NESOI, SPARK IGN,4 CYL, 1500-3000CC				
(HS: 8703230044) (SIC: 3711)				
Unit of Quantity -- Number				
1994	361,535	4,606,893,087	1,169,719	6,474,810,449
1995	314,175	3,963,197,536	1,125,534	5,396,909,612
1996	322,248	4,381,403,774	1,295,532	5,630,167,625
1997	488,692	6,661,333,393	1,184,855	7,458,690,130
1998	474,633	6,091,373,316	912,740	6,573,997,869

Figure 1: Import Data for Automobiles

along with the classifications which have been manually assigned to the document.

The final stage to creation of an Entry Vocabulary Index is to develop a maximum likelihood weighting associated with each term (word or phrase) and each classification. One constructs a two-way contingency table for each pair of word/phrase terms t and classifications C as shown in table 1. where a is the number of document titles/abstracts

	C	$\neg C$
t	a	b
$\neg t$	c	d

Table 1: Contingency table from words/phrases to classification

containing the word or phrase and classified by the classification; b is the number of document titles/abstracts containing the word or phrase but not the classified by the classification; c is the number of titles/abstracts not containing the word or phrase but is classified by the classification; and d is the number of document titles/abstracts neither containing the word or phrase nor being classified by the classification.

The association score between a word/phrase t and an classification C is computed following Dunning [4]

$$\begin{aligned} W(C,t) &= 2[logL(p_1,a,a+b) + logL(p_2,c,c+d) - \quad (1) \\ &= logL(p,a,a+b) - logL(p,c,c+d)] \quad (2) \end{aligned}$$

where

$$logL(p,n,k) = klog(p) + (n-k)log(1-p) \quad (3)$$

and $p_1 = \frac{a}{a+b}$, $p_2 = \frac{c}{c+d}$, and $p = \frac{a+c}{a+b+c+d}$.

3. APPLICATIONS

3.1 Cross-language search

A very interesting application of Entry Vocabulary Indexes is to multilingual information access. Because large university electronic catalogs contain bibliographic references for thousands of documents in foreign languages (the Library of Congress language list contains 400 languages), one can build EVIs which map to the (English) Library of Congress Subject Headings (LCSH). Library catalogers typically manually index and assign multiple LCSH entries to each book or other item being cataloged. Our training set for construction of a multilingual EVI for LCSH is the six million record set of the University of California MELVYL online catalog (http://www.melvyl.ucop.edu). As the following figure demonstrates, one can enter foreign language words and be pointed to the

subject headings which most closely match on a maximum likelihood basis. This subject heading can be used as a reliable search query in online library catalogs, since LCSH is an industry standard. In the example, the German query word "Wirtschaftspolitik" presents the subject heading "Economic Policy" as its top ranked metadata item. This happens to be an exact translation of Wirtschaftspolitik.

Our initial use of EVIs has been applied to cross-language search of the NTCIR collection of Japanese-English scientific documents [5] and more recently to English-German retrieval for the domain specific task of the CLEF European language evaluation [7] on the GIRT collection of German documents in the social science domain.

3.2 Numeric data

The example of import data in the introduction demonstrates an important genre of digital objects for which search is difficult. Numeric statistical databases, their classifications and descriptions could be called 'evidence poor' because they lack the rich and abundant textual clues so important in information discovery. Neither string search (as provided by the sites) nor inverted word indexing will properly search the data. Yet the humanly indexed categories within each classification scheme contain a precise description of that category, useable if you are expertly knowledgeable about the details of foreign trade. To provide search support for novice or non-expert searching, we must somehow expand the search possibilities.

We can do this by mining the textual resources of electronic libraries in much the same way as above for cross-language search. A large selection of trade magazine abstracts in these libraries have been indexed manually by the assignment of the very same category codes used to classify the numeric statistical information. For example a magazine article about the new management directions of Apple

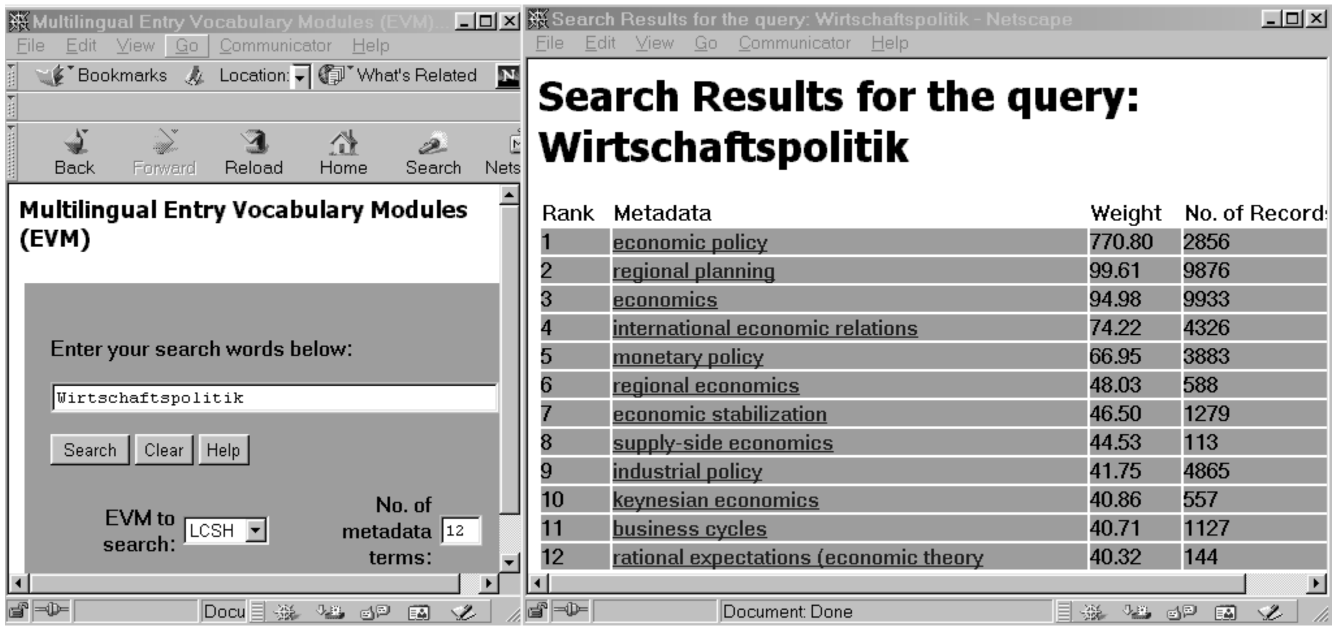

Figure 2: EVI Multilingual Search against Library of Congress Subject Headings

Computer will be assigned (by a human indexer) the industrial classification code 3571 for 'Electronic Computers'. By mining the textual descriptions found in the documents (either titles or abstracts), one can build an Entry Vocabulary Indexes which map ordinary words and phrases to the classification codes. Examples of such entry vocabulary modules can be found at the site
http://www.sims.berkeley.edu/research/metadata
under the 'prototypes' section. The one for 'SIC' will demonstrate entry vocabulary search for U.S. Imports and Exports. The process by which this numeric classification EVI was constructed has been described in [6, 8].

More recently we have taken the 1997 Economic Census for which the Census Bureau provides a selection and display mechanism http://www.census.gov/epcd/www/econ97.html for data summarized to the North American Industrial Classification (NAICS) coding system [10]. The census system lacks the specificity to address particular instances of companies associated with NAICS codes. However, our NAICS EVI prototype (at the url above) will take the query 'IBM' and return a selection of NAICS codes (by entry vocabulary mapping from textual documents indexed by these codes) of industries closely associated with IBM's corporate activities (see Figure 3).

3.3 Patents and Other Specialty Areas

Multiple Entry Vocabulary Indexes have been built for the U.S. Patent Databases. The documents in the U.S. Patent office system have been indexed by both the U.S. patent classification system and the international Patent classification system of the World Intellectual Patent Organization (WIPO). Other EVIs were constructed for the INSPEC service (science and engineering abstracts) and MEDLINE (medical specialties).

4. EVALUATION STRATEGIES

Since EVI technology and prototypes have only been available for the past year or so, formal evaluation has yet to be undertaken. The DARPA Information Management program is funding an in-depth evaluation of this technology with one or more of the following evaluation strategies:

- TREC-like recall precision improvement for specific tasks
- Hands-on interactive search with/without EVI
- Web session log analysis

Each of these strategies could be used to test search with or without the use of an entry vocabulary module as if they were two different systems. We have performed preliminary TREC-style evaluations for cross-language conferences and they show promising improvements over retrieval without EVIs.

5. CONCLUSIONS AND FUTURE WORK

5.1 Summary

Entry vocabulary technology, in the form of Entry Vocabulary Indexes, offers a new approach to digital object search. The approach capitalizes and leverages the worldwide investment in human indexing and development of manual classification schemes, subject indexes and thesauri. Its central feature incorporates a probabilistic mapping between ordinary language and technical vocabulary or classifications. The technology may be applied to digital genres not normally associated with textual search, such as numeric statistical databases. A more detailed discussion of vocabulary, metadata and search may be found in [3].

5.2 Search with non-Roman Scripts

For the future, we are interested in dealing with languages with other than a latin or Roman alphabet. Consider, for

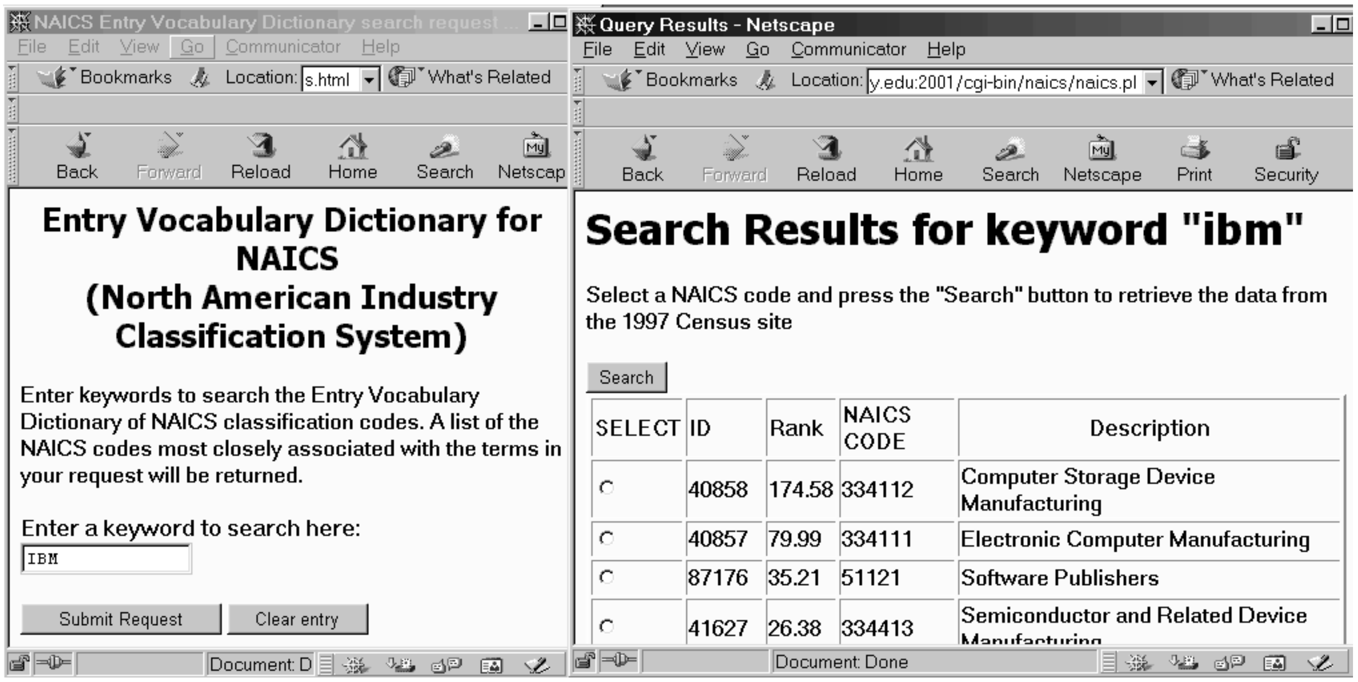

Figure 3: NAICS Search Example

example the Russian phrase економическаиа политика. The electronic libraries of the United States follow a cataloging standard [1] for transformation (called 'transliteration' or 'Romanization') from non-Roman scripts into their Romanized equivalents. For example, the Cyrillic letter щ is expressed as 'shch'. Such transformations are one-to-one and reversible. If we prepared a transliteration front end to the above Cyrillic input, it would obtain the phrase "ekonomicheskaia politika." Submitting this phrase to the Entry Vocabulary Index for the Library of Congress subject headings, it should return the same subject heading "economic policy" as the previous German term "Wirtschaftspolitik." We are in the process of developing such transliteration and EVI search for the Cyrillic alphabet.

5.3 Images

Image data provides an interesting challenge for EVI technology. In work conducted in conjunction with the NSF/NASA/DARPA-sponsored Digital Library Initiative project [9], "blob" representations (each "blob" is a coherent region of color and texture within an image) were derived from a collection of 35000 images and indexes were created for probabilistic matching of images, based on representations of the blobs. Since each of the images in the "BlobWorld" database have associated keywords in their metadata records, we are able to apply the same basic EVI concept to these image records.

In this case the metadata keywords describing the images are associated with the individual blobs extracted from the images. Thus we are building a probabilistic association between certain keywords and patterns of color and texture in the image database. For example blobs with orange and black stripes might be associated with the keyword "TIGER".

6. ACKNOWLEDGMENTS

Entry Vocabulary Technology has been developed under support by Defense Advanced Research Projects Agency (DARPA) Information Management program through DARPA Contract N66001-97-8541; AO# F477: Search Support for Unfamiliar Metadata Vocabularies. Application of the EVI technology to cross-language retrieval was supported by research grant number N66001-00-1-8911 (Mar 2000-Feb 2003) from the DARPA Translingual Information Detection Extraction and Summarization (TIDES) program. Application of EVI technology to numeric data search was supported by a National Library Leadership award from the Institute of Museum and Library Services entitled "Seamless Searching of Numeric and Textual Resources."

Many graduate students have been associated with the development of various phases of the entry vocabulary technology, chief among them Barbara Norgard and Youngin Kim. Other contributions were made by Hui-Min Chen, Michael Gebbie, Natalia Perelman, Vivien Petras, and Jacek Purat.

7. REFERENCES

[1] Randall K. Barry. *ALA-LC romanization tables : transliteration schemes for non-Roman scripts.* Washington : Cataloging Distribution Service, Library of Congress, 1997.

[2] Eric Brill. A simple rule-based part of speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*, 1992.

[3] Michael Buckland et al. Mapping entry vocabulary to unfamiliar metadata vocabularies. In *D-Lib Magazine.* http://www.dlib.org/dlib/january99/buckland/01buckland.html, January 1999.

[4] Ted Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*,

19:61–74, March 1993.

[5] Fredric Gey, Aitao Chen, and Hailing Jiang. Applying text categorization to vocabulary enhancement for japanese-english cross-language information retrieval. In S. Annandiou, editor, *The Seventh Machine Translation Summit, Workshop on MT for Cross-language Information Retrieval, Singapore*, pages 35–40, September 1999.

[6] Fredric Gey et al. Advanced search technologies for unfamiliar metadata. In *Proceedings of the Third IEEE Metadata Conference*. IEEE, 1999.

[7] Fredric Gey, Hailing Jiang, Vivien Petras, and Aitao Chen. Cross-language retrieval for the clef collections - comparing multiple methods of retrieval. In Carol Peters, editor, *Proceedings of the CLEF 2000 Cross-Language Text Retreival System Evaluation Campaign*. Springer, to appear 2001.

[8] Youngin Kim, Barbara Norgard, Aitao Chen, and Fredric Gey. Using ordinary language to access metadata of diverse types of information resources: Trade classification and numeric data. In *Proceedings of the 62nd Annual Meeting of the American Society for Information Science*, pages 172–180. ASIS, 1999.

[9] Ray R. Larson and Chad Carson. Information access for a digital library: Cheshire ii and the berkeley environmental digital library. In *Proceedings of the 62nd Annual Meeting of the American Society for Information Science*, pages 515–535, November 1999.

[10] U.S. Office of Management and Budget. *North American Industry Classification System*. Maryland: Berman Press, ISBN 0-89059-09704, 1997.