Stories

D-Lib Magazine
January 1999

Volume 5 Number 1

ISSN 1082-9873

Mapping Entry Vocabulary to Unfamiliar Metadata Vocabularies

Michael Buckland, with Aitao Chen, Hui-Min Chen, Youngin Kim, Byron Lam, Ray Larson, Barbara Norgard, and Jacek Purat
(buckland, aitao, hmchen, kimy, byronlam, ray, barbara, )
School of Information Management & Systems;
and Fredric Gey, ()
UCData

University of California
Berkeley, CA 94720

http://www.sims.berkeley.edu/~buckland/

Introduction

The emerging network environment brings access to an increasing population of heterogeneous repositories. Inevitably, these, have quite diverse metadata vocabularies (categorization codes, classification numbers, index and thesaurus terms). So, necessarily, the number of metadata vocabularies that are accessible but unfamiliar for any individual searcher is increasing steeply. When an unfamiliar metadata vocabulary is encountered, how is a searcher to know which codes or terms will lead to what is wanted? This paper reports work at the University of California, Berkeley, on the design and development of English language indexes to metadata vocabularies. Further details and the current status of the work can be found at the project website http://www.sims.berkeley.edu/research/metadata/

The Significance of Unfamiliar Metadata

Many of the most information-rich repositories, especially bibliographical and textual databases, have some form of categorization, classification, coding, or indexing. An increasing number of techniques are being developed for automatic categorization of repositories for which human indexing is unavailable. Experienced searchers know that familiarity with the source being searched, whether database or reference work, is critical for effective, reliable searching. Each source has its own quirks and personality and familiarity comes from experience, from frequent use. Indeed, the more that has been invested in the enhancement of the source, the richer the metadata and the more important this personal experience and familiarity become, and the less they can be used effectively or efficiently except by searchers who are familiar with them.

There is a massive investment world-wide in making repositories accessible over networks and a major investment world-wide in providing indexing, categorizing, and other metadata. So the number and proportion of network accessible repositories with unfamiliar metadata vocabularies are rapidly growing. The amount of searching can be expected to rise, but diminishing search effectiveness is the predictable result. Meanwhile libraries are working to provide support for

access to this wider world [Norgard et al. 1993].

The explosive increase in heterogeneity assures that the lack of familiarity required for efficient, effective searching is an increasing problem. When an index or categorization scheme is encountered, how is one to know what word or value has been assigned to the topic that one is interested in? Expert human search assistance is often needed, but a sufficient population of human expert search intermediaries is unaffordable. The challenge, therefore, is to provide automatically the kind of expert prompting that an expert human search intermediary would provide. It has been argued that the most cost-effective single investment for improving effectiveness in the searching of repositories would be technology to assist the searcher in coping with unfamiliar metadata vocabularies [Buckland 1992].

Examples

An information system which utilizes a specialized vocabulary for classification and search purposes is the Census Bureau's U.S. Imports and Exports numeric data issued on CD-ROM and accessible at http://govinfo.kerr.orst.edu/impexp.html. These data are important for people engaged in strategic policy and investment decisions. Imagine, for example, that data relating to the auto industry were required. A commodity search using the term "automobile" will find nothing. A search on "cars" will lead only to "Railway or Tramway Stock." Yet the data are there... under "Passenger Motor Vehicles, Spark Ignition Engine."

Rockets are increasingly used in military hostilities. How many does the U.S. export? A search of the exports data using the word "Rockets" yields the commodity category "Bearings, Transmission, Gaskets, Misc." Restricting the search to the singular form "rocket" yields an additional three categories:

> Photographic or Cinematographic Goods
> Engines, Parts, Etc
> Arms and Ammunition, Parts and Accessories Thereof

The last of these specifically concerns military weapons exports from the United States. Indeed the specific term "rocket" is found only in the category MISSILE & ROCKET LAUNCHERS AND SIMILAR PROJECTORS (9301009050) and completely misses a larger export category GUIDED MISSILES (9306900020), while the general heading category for this section is: BOMBS, GRENADES, ETC (9306). Clearly researchers who wish to mine this database need a tool that will bridge the gap between common terminology and the highly specialized classification scheme which has evolved for categorizing these data.

Sometimes the effects are subtle as well as unexpected, as in this example of exact subject searches in the University of California's MELVYL online library catalog:

> FIND XSU VIETNAM WAR
> Search result: 0 records
>
> FIND XSU VIETNAMESE CONFLICT
> Search result: 4,190 records

Berger's detailed analysis of search sessions in the MELVYL catalog has shown that library users are quite adept at recovering from errors when searching for names, but that they are significantly less able to recover from errors when searching subject headings [Berger 1994].

A Remedy: Indexes to Metadata Vocabularies

The Dewey Decimal Classification number "330" denoting Economics is a kind of a word with an odd appearance. Dewey numbers convey meaning, if you are familiar with the numbers, but the meanings are more or less unclear until familiarity is developed through usage. What helps is a mapping from our English terms to the Dewey Decimal Classification numerical terms, an

English to "Dewey" dictionary. Melvil Dewey provided this in the form of his *Relativ Index*. We tend to take this Relativ Index for granted as a necessary ancillary to the classification, but Dewey himself considered it the most important part of his system. Using his reformed spelling, Dewey wrote:

"This alfabetic Index, the most important feature of the sistem, consists of headings gatherd from a great variety of sources, as uzers of the sistem hav found them desirabl.... The Index givs similar or sinonimus words,... so any intelijent person wil surely get the ryt number.... The Relativ Index, with its cachwords... insures that books on same faze of any subject cuming before the clasifyers shal be assynd to same place, and that any reader seeking these books shal be referd instantly to that place." [Olding 1966, 82-91]

The need for an English language index to the Dewey Decimal Classification dictionary may seem obvious. What is less obvious is the need for a comparable dictionary in less exotic cases where so-called natural language is used, often adapted in rather unnatural ways. The purpose is to provide guidance in the transition from familiar vocabulary, whether ordinary English or from a familiar categorization scheme, to the unfamiliar, providing the kind of expert prompting that an expert search intermediary would. In our work we call this kind of search aid an Entry Vocabulary Module. The Entry Vocabulary Module helps the searcher to be more effective and, thereby, provides a value-added enhancement, increasing the return on the original investment in generating metadata.

Prototype Entry Vocabulary Modules

Our research in this field has the following main threads:

a. Development of tools to support the creation of Entry Vocabulary Modules;

b. Creation of a set of prototype Entry Vocabulary Modules for a challenging range of examples, including subdomains within databases;

c. Deployment as an amenity provided by a repository; as a network-accessible amenity; and embedded in a individual or collaborative work environment;

d. The use of natural language processing techniques in addition to statistical term co-occurrence; and

e. Recommendations for the improvement of metadata documentation for numeric databases.

Figure 1 is an overview of the current infrastructure design.

Figure 1

The prototype Entry Vocabulary Modules that we have developed are web-accessible at <http://www.sims.berkeley.edu/research/metadata/oasis.html>.

They include English language indexes to BIOSIS Concept Codes, to the INSPEC Thesaurus, and to the U.S. Patent and Trademark Office Patent Classification, and a multilingual index (supporting queries in English, French, German, Russian, or Spanish) to the physical sciences sections of the Library of Congress Classification,. When the Entry Vocabulary Module leads to a promising term in the target metadata vocabulary, a search can then be executed using the newly-found metadata in a remote database. Because of licensing restrictions, however, extending the search to a remote database is restricted to the Patent and Library of Congress classifications, unless the searcher has a Berkeley IP address. An English language index to the Standard Industrial Classification (SIC) codes has recently been added and will be linked to numeric database.

Research Aspects

The remainder of this article summarizes the current state of our research on Entry Vocabulary Modules.

Statistical Association

Examples already exist of assistance with the selection of search terms. These are ordinarily limited to the terms already present in the target metadata. They do not, except incidentally, enable other, ordinary English discourse to be used as a point of entry. An example is the Grateful Med, which provides a front end to the MeSH headings for subject access to medical literature.

The outstanding example of providing links between vocabularies is the National Library of Medicine's Unified Medical Language System (UMLS), which depends on expensive, intensive human expertise in establishing the links. The work reported here offers a different and complementary approach. We intend a very low-cost, computer-generated alternative for domains and metadata outside those in the health field covered by the UMLS.

Our 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]. He used probabilistic interpretation of vector-spaced retrieval, extended by William Cooper's Staged Logistic Regression 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].

Natural Language Techniques

In Larson's original method, Entry Vocabulary Modules were derived from the frequency of occurrence of single terms. More recently natural language parsing software has been used to identify noun phrases and to use these phrases instead of the individual words within them [Kim & Norgard 1998].

Subdomains

It is obvious that different languages, such as Chinese, English, and German, use different words. Also, within any given language, different domains use differing vocabularies. These differences are often more extensive than expected. Suppose a searcher looked in the *Library of Congress Subject Headings* and in the *Medical Subject Headings* (MeSH) for "Coastal pollution." Neither system uses that phrase, but, instead, in ranked order:

LCSH: Marine pollution; Coastal zone management; Water -- Pollution; Petroleum industry and trade; Beach erosion; Coasts; Barrier islands; Coastal changes; etc.

MeSH: Seawater; Water pollution; Bacteria; Water microbiology; Air pollution; Environmental monitoring; Bathing beaches; Environmental pollution; etc.

Note how different the two lists are and note the variety within each. It is easier to for a person to recognize pertinent terms than to predict them. Each of these subject headings is individually plausible, but who could be expected to imagine them? Three different vocabularies are simultaneously in use here: LCSH, MeSH, and the searcher's.

Ordinarily metadata vocabularies are studied as a whole, but, even within a database, searchers are rarely equally interested in all parts of a database for any given search. They are usually interested in some particular subdomain. Vocabulary can vary for subdomains even within a single database, suggesting a need for separate Entry Vocabulary Modules for topical

subdomains. As an example we have developed three subdomain Entry Vocabulary Modules in the INSPEC thesaurus: for "Biology", for "Information Science," and for "Water." If the same search terms are given to these variant subdomain Entry Vocabulary Modules, they can be expected to respond with different suggested thesaural terms because different semantic contexts are assumed. The sensitivity of actual search results to these differences has yet to be examined [Kim 1998].

To obtain a training set for a given subdomain one can use journals that rank highly in that area domain as sources of representative data. Each year the Institute for Scientific Information publishes a report ranking journal impact by measuring the number of times each journal article was cited. We have used rankings from the *Science Citation Index Journal Impact Report* (SCI) and the *Social Science Citation Index Journal Impact Report* (SSCI) to select journal titles representative of a particular domain.

Numeric Databases and "Portable" Indexes

In databases of text, "full-text" searching for words in the database content provides an alternative to searching the metadata. This is impractical with databases with non-textual content, such as numeric databases, where the metadata (the "codebooks") tend to be meager.

However, some metadata vocabularies, such as the Standard Industrial Classification (SIC), are used in both text and non-text databases. It may be possible to create, say, an English to SIC Entry Vocabulary Module from a text database with SIC codes. The English to SIC index could then also be used with a numeric database with SIC codes. The SIC is a good example because it is extensively used in numeric databases for which an Entry Vocabulary Module could not be generated directly. We have generated an Entry Vocabulary Module from a bibliographic database and intend shortly to link it to a numeric database.

Creating Metadata

An Entry Vocabulary Module provides a ranked list of probably relevant metadata terms for any fragment of text, so it can also be used for computer-assisted categorization if text is submitted as if it were a search query. Larson's "classification clustering" methodology, upon which our work is based, was used to assign Library of Congress Classification numbers to 283 new books. For 47% of the records the first-ranking number selected by automatic classification was the same class number as the one actually assigned in a local library. In many other cases, the first-ranked number was a plausible alternative and the top ten numbers included the number humanly assigned. [Larson 1992].

Implementation Contexts

Because there are so many different metadata systems and because different groups of searchers bring the vocabulary of their own language and specialty, Entry Vocabulary Modules have very extensive implementation potential. Further, they can be usefully positioned in several ways: as an amenity on the searcher's client to provide assistance when accessing an unfamiliar remote repository; as an amenity on (or invoked from) a repository server to aid remote searchers unfamiliar with the local metadata scheme of that repository; or as an amenity on a work-centered computing environment.

Future Work

A set of prototype Entry Vocabulary Modules is being developed for a diverse selection of challenging metadata environments. Intelligent agent software capable of supporting the creation of new Entry Vocabulary Modules for remote repositories is under development [Norgard, 1998]. The sensitivity of Entry Vocabulary Modules to variations between subdomains within repositories is to be examined. We hope, in the light of our experience, to be able to make recommendations for the improvement of "codebook" metadata documentation for numeric databases.

Summary

This work represents a confluence of three lines of research: The OASIS program of studies in adaptive searching, led by Michael Buckland, which from 1990 has developed prototypes for supporting improved use of existing metadata [Buckland et al. 1992, and http://www.sims.berkeley.edu/research/oasis/]; Ray Larson's development of "classification clustering" to create entry vocabulary modules for the Library of Congress Classification in CHESHIRE, a next-generation online catalog and full-text information retrieval system using advanced IR techniques [http://cheshire.lib.berkeley.edu/]; and Fredric Gey's research and development on access to numeric databases and use of probabilistic retrieval techniques in TREC [http://ucdata.berkeley.edu/gey.html].

The rapid increase in network-accessible repositories increases opportunities for searching. Unfamiliar metadata is difficult to search and, increasingly, what is accessible is unfamiliar. Providing indexes to metadata vocabularies offers a remedy.

We seek to exploit the potential significance of combining linguistic analysis with statistical methods to help searchers. We are designing an amenity that can easily be used on regular workstations and integrated into actual work environments with little investment of time and effort in order to improve performance when searching and thereby to improve the return on the investment in those systems.

Advantages of this work are that it provides an alternative to the expensive human crafting of links within and between vocabularies, it is based on searching of fragments existing within the metadata and databases, it uses advanced probabilistic techniques, it allows the searcher to start a search using familiar search terms from a familiar vocabulary, it is designed for use in a networked environment, and, since only a training set is required, it allows for rapid deployment with unfamiliar metadata schemes.

Acknowledgment

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

The development of Entry Vocabulary Modules was part of the OASIS research program developing prototypes of enhancements to online library catalogs and other databases. For OASIS, see [Buckland et al. 1992] and http://www.sims.berkeley.edu/research/oasis/

OASIS research has been supported by the US Department of Education (HEA IIA) and by the Berkeley Digital Libraries Initiative project NSF IRI-9411334.

References

[Berger 1994] Michael G. Berger. 1994. *Information-Seeking in the Online Bibliographic System: An Exploratory Study*. Ph. D dissertation, University of California, Berkeley, School of Library and Information Studies. UMI #AAI9504745. Detailed analyses of how searches progress during a search session.

[Buckland 1992] Buckland, M. K. 1992. Agenda for online catalog designers. *Information Technology and Libraries* 11, no. 2 (June 1992):157-163. On the strategic significance of providing support for searching unfamiliar metadata vocabularies.

[Buckland et al. 1992] Buckland, M. K., M. H. Butler, B. A. Norgard & C. Plaunt. 1992. OASIS: A Front-End for Prototyping Catalog Enhancements. *Library Hi Tech*, Issue 40 (1992):7-22. Summary of prototyped information search aids.

[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. Demonstration of the fundamental algorithms of probabilistic text retrieval using logistic regression as the machine learning technique.

[Kim 1998] Kim, Youngin. 1998. *Sensitivity of Entry Vocabulary Modules to Subdomains*. Technical report.
http://www.sims.berkeley.edu/research/metadata/subdomain.html

[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.

[Larson 1992] Larson, R. R. 1992. Experiments in Automatic Library of Congress Classification. *Journal of the American Society for Information Science*, v. 43 no. 2 (March 1992), pp. 130-148

[Norgard 1998] Norgard, Barbara. 1998. *Entry Vocabulary Modules and Agents*. Technical report.
http://www.sims.berkeley.edu/research/metadata/agents.html

[Norgard et al. 1993] Norgard, B. A., M.G. Berger, M. K. Buckland, & C. Plaunt. 1993. The Online Catalog: From Technical Services to Access Service. *Advances in Librarianship* 17 (1993):111-148. Extensive, evaluative review of literature on problems and innovation on online bibliographic retrieval systems.

[Olding 1966] Olding, R. K., ed. 1966. *Readings in Library Cataloguing*. Hamden, CT: Archon Press.

[Plaunt & Norgard 1998] 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 (August 1998): 888-902.
<http://bliss.berkeley.edu/papers/assoc/assoc.html>

Top | Contents
Search | Author Index | Title Index | Monthly Issues
Previous Story | Next Story
Comments | Home| E-mail the Editor

D-Lib Magazine Access Terms and Conditions

DOI: 10.1045/january99-buckland