Domain-Based Indexes:
Indexing for Communities of Users

Michael Buckland[1], Hailing Jiang[1], Youngin Kim[1], Vivien Petras[2]

[1] *School of Information Management and Systems*
University of California, Berkeley, USA 94720-4600
{buckland, hjiang1, kimy@sims.berkeley.edu}

[2] *Institut fuer Bibliothekswissenschaft, Humboldt-Universitaet zu Berlin,*
Unter den Linden 6, 10099 Berlin, Germany
vivien.petras@rz.hu-berlin.de

Résumé :

La formation d'un vocabulaire évolue à la foi au sein d'une communauté et d'un domaine discursif. Cependant, les bases de données bibliographiques ont souvent un seul index créé pour la base entière, et ceci bien qu'elles couvrent fréquemment plusieurs domaines discursifs. A des fins expérimentales, des indexes furent dérivés du langage utilisé au sein d'un domaine discursif specialisé, sous-ensemble d'une base de données. Ce radical éloignement des pratiques traditionelles produit une amélioration significative des performances de recherche. La conclusion que les performances sont meilleures au sein d'un domaine spécifique, et qu'elles se détériorent au fur et à mesure que la portée du système s'étend a des domaines additionnels, est conforme aux expériences conduites en intelligence artificielle et traduction par machine. Cette analyse a également nécessité le développement d'une mesure opérationnelle de la performance des intermédiaires. Ceci résulte en plusieurs questions théoriques et pratiques.

Mots-clés : Indexation, vocabulaire, domaines discursifs, meta-données.

Abstract :

The formation of vocabulary evolves within communities, within domains of discourse. However, bibliographic databases have traditionally had one single index created for the entire database, even though bibliographic databases usually cover an arbitrary group of domains of discourse. As an experiment, indexes derived from the language used within one single specialized domain of discourse, a subset of a database. This radical departure from traditional practice shows significant improvements in retrieval performance. The conclusion that performance is best within specific domains and deteriorates as the scope of the system expands to include additional domains is consistent with experience in artificial intelligence and in machine translation. Analysis has required the development of an operational measure of the performance of intermediaries. Several theoretical and practical questions arise.

Keywords : Indexing, vocabulary, domains, metadata.

1. Collection-based indexes

Collections of documents, such as bibliographies, catalogs, or collections of images or texts, commonly have topical (or "subject") indexes. Frequently, verbal indexes -- subject headings or thesauri -- are used. The significance of verbal subject indexes extends beyond lists of subject headings and thesauri. They are also needed to enable use of classification systems and other non-verbal categorization systems. An example is the Relative Index to the Dewey Decimal Classification. Even experienced searchers need a subject index -- in words -- to identify what the appropriate classification number would be. Melvil Dewey considered the Relative index to his classification to be at least as important as the classification itself.

In this paper we are concerned with Relative Indexes, also known as Entry Vocabulary Indexes, which provide an index (or mapping, or bilingual dictionary) from the words with which searcher might begin a search ("Query Vocabulary") and the terms in the formal, system metadata, such as the INSPEC Thesaurus ("Entry Vocabualry"). Examples can be accessed and used at www.sims.berkeley.edu/research/projects/metadata/GrantSupported/seamless_prototypesI.html.

2. Communities of Discourse

The vocabulary of natural languages evolves distinctively within communities. "Dialects" of word-usage evolve because specialized meanings develop through metaphor for particular purposes, new words are coined, and phrases of local significance evolve. Meaning depends on context. Every community has its distinctive vocabulary and, indeed, each community is characterized by, and can be identified by, its vocabulary.

Previous research in information science has been aware of differences in vocabulary between different domains of discourse. Birger Hjorland's *Information Seeking and Subject Representation: An Activity-Theoretical Approach to Information Science* (Greenwood, 1997) is a noteworthy example [HJORLAND 97]. However, discussion has ordinarily been in terms of differences between broad "disciplines" and in the well-known differences in vocabulary (and therefore, subject indexes) between different discipline-based databases, such as *Chemical Abstracts*, *INSPEC*, and *Medline*, and also between each of these disciplines and universal subject indexes used to cover all topics, such as the *Library of Congress Subject Headings*.

Normal practice has been to create “collection-based” indexes. That is to say that the index is to the collection (database, repository) as a whole. This is the obvious course of action and, until recently, we are aware of no exceptions to this practice.

3. Domains of Specialized Discourse

The reality is that even specialized, discipline-based databases are not internally homogeneous in their use of vocabulary. The scope of databases such as *INSPEC* or *Medline* are in reality defined by an arbitrary, albeit judicious, boundary drawn around a group of related subdomains. But each individual subdomain has its own vocabulary, its own distinctive terminological practices. In our research at the University of California, Berkeley, we have been concerned with how to make indexes both easier to use and also more effective. Recently, we have examined the consequences of creating indexes based on individual, small domains with specialized discourse instead of the totality of the collection being indexed.

What can be expected to follow from this different basis is that an index based on the word-usage of a single (sub)domain is likely to be more satisfactory, to perform better, for searchers and searches within that subdomain. Preliminary evaluation of subdomain indexes show this to be markedly true.

4. On the Evaluation of Indexes

We use the phrase Entry Vocabulary Index to denote a mapping from Query Vocabulary to Entry Vocabulary. The technique employed is to use the terminology in titles and abstracts as a surrogate for Query Vocabulary and then use statistical techniques to indicate the degree to which each word and phrase in the titles and abstracts is associated with each individual metadata value (Entry Vocabulary) [BUCKLAND 99 ; PLAUNT 98].

Evaluation requires a methodology for measuring the performance of such indexes. We used a test developed by Larson [LARSON 92]: If titles of new documents are used as queries, can the index predict what index terms will have been assigned to them by the database indexer? Further, formal measurement in terms of Precision and Recall can be adopted: Instead of predicting the performance of an information retrieval system in selecting (retrieving) relevant documents, an entry vocabulary index can be judged by how well it identifies (predicts) the “relevant” subject index terms where the terms assigned by the indexer are considered to be the “relevant” terms [KIM 00]. This process can be considered to be a methodology for evaluating the performance of intermediaries.

A test was performed using queries (titles) within the domain of Astrophysics in INSPEC, a bibliographical, abstracting services covering the literature of computing, engineering and physics. The results show that an entry vocabulary index based on the vocabulary of Physics performed significantly better than an entry vocabulary index based on the entire database, and that an entry vocabulary index derived (only) from the discourse on Astrophysics performed significantly better than that derived from the literature of Physics [BUCKLAND 00].

In a second experiment, the INSPEC classification scheme was divided into thirty-one subdomains. Thirty-one separate entry vocabulary indexes were created, one from a sample of records in a single subdomain, and also one general entry vocabulary index from a sample drawn form the entire database. Then a sample of titles was collected from the subdomains and submitted as a query three times:

1. To the general Entry Vocabulary Index derived from the entire database;
2. To its "own" Specialized Entry Vocabulary Index, meaning the Index for the subdomain from which the title had been taken; and
3. To a specialized Specialized Entry Vocabulary Index selected at random and so, probably not its "own" subdomain.

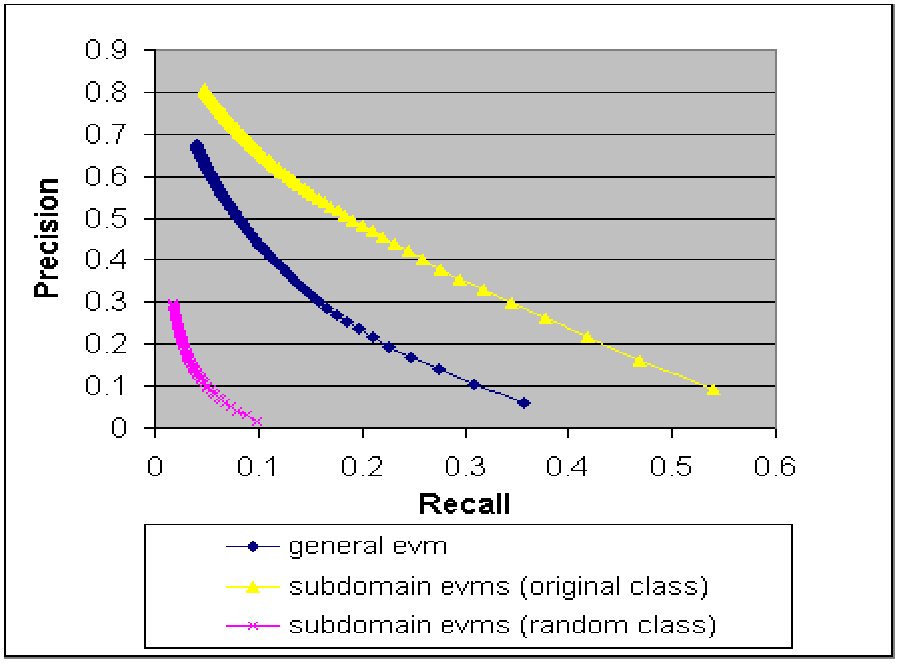

Figure 1: Sensitivity of Performance to Choice of Index.

The results, shown in Figure 1, indicate that submitting a query to a specialized entry vocabulary index based on discourse used in the specialized subdomain of the query significantly improves search performance. Using a general entry vocabulary index based on the entire database is less effective. But using multiple specialized indexes has its risks. A query submitted to a specialized entry vocabulary index based on the discourse of a different subdomain performs badly compared with a general entry vocabulary index.

5. Discussion

Several theoretical and practical problems arise:

1. How should we identity and delineate domains of discourse?
2. How narrow or wide should the domain selected be?
3. The smaller the domain the smaller the basis (training set) for creating entry vocabulary indexes and, perhaps, the better the search performance. But the smaller the basis for the sample the fewer the range of words and phrases included and the narrower the capability of the index.
4. Will we find the same situation in the social science and in the humanities?
5. If we use specialized entry vocabulary indexes, how can we choose the correct index?
6. How stable are specialized vocabularies over time?

The creation of multiple, different indexes for the same database for different specialized domains was not economically feasible until the arrival of digital databases and automatic algorithms for generating indexes. Such specialized indexes promise significant improvements in service but important questions remain to be investigated.

6. Acknowledgements

The work reported here is part of a larger program directed by Fredric Gey, Ray R. Larson, and Michael Buckland. It has been supported by the Institute of Museum and Library Services and by the Defense Advanced Research Projects Agency. Further details can be found at www.sims.berkeley.edu/research/projects/metadata/

7. References

[BUCKLAND 99] BUCKLAND AND OTHERS. “Mapping Entry Vocabulary to Unfamiliar Metadata Vocabularies.” *D-Lib Magazine*, vol. 5, no. 1, Jan 1999. www.dlib.org/dlib/january99/buckland/01buckland.html

[BUCKLAND 00] BUCKLAND AND OTHERS. “Variation by Subdomain in Indexes to Knowledge Organization Systems.” In: *Dynamism and Stability in Knowledge Organization: Proceedings of the Sixth International ISKO Conference, 10-13 July 2000, Toronto, Canada*. Ed. By Clare Beghtol, L. C. Howarth, N. J. Williamson. Wuerzburg, Germany: Ergon Verlag, 2000. Pp. 48-53. www.sims.berkeley.edu/research/metadata/iskopaper.html

[HJORLAND 97] HJORLAND, BIRGER. *Information Seeking and Subject Representation: An Activity-Theoretical Approach to Information Science*. Greenwood Press, Westport, CT.

[KIM 2000] KIM, YOUNGIN. Evaluation of the performance of the EVM dictionaries. School of Information Management and Systems, University of California, Berkeley, CA, USA 94720-4600. June 30, 2000
http://www.sims.berkeley.edu/research/projects/metadata/ResearchAreas/EvalMethods.htm

[LARSON 92] LARSON, RAY. R. "Experiments in Automatic Library of Congress Classification." *Journal of the American Society for Information Science*, vol. 43 no. 2 (March 1992), pp. 130-148.

[PLAUNT 98] PLAUNT, C., AND B. A. NORGARD. "An Association Based Method for Automatic Indexing with a Controlled Vocabulary." *Journal of the American Society for Information Science*. vol. 49, no. 10, August 1998, pp. 888-902.
www.sims.berkeley.edu/research/metadata/assoc/assoc.html