

The Effectiveness of Thesauri-Aided Retrieval

Kazem Taghva*, Julie Borsack, and Allen Condit

Technical Report 98-01

Information Science Research Institute

University of Nevada, Las Vegas

June 1998

Abstract

In this report, we describe the results of an experiment designed to measure the effects of automatic query expansion on retrieval effectiveness. In particular, we used a collection-specific thesaurus to expand the query by adding synonyms of the searched terms. Our preliminary results show no significant gain in average precision and recall.

*Email: taghva@cs.unlv.edu

1 Introduction

The notion of retrieval effectiveness is concerned with the ability of the system to retrieve relevant documents while at the same time suppressing the retrieval of non-relevant documents. It is this notion that has led to the discovery of three prominent models, or different approaches to building, retrieval systems. These models are the Boolean model, the Vector Space model, and the Probabilistic model[6].

In conjunction with these models, developers have designed many tools to enhance retrieval effectiveness. Two prominent tools are feedback methods and thesaurus-aided retrieval. In this report, we concentrate on the effect of thesauri on retrieval in BASISplus, a system based on Boolean logic.

A thesaurus provides a precise and controlled dictionary which can be used to coordinate document indexing and retrieval. It can also be used to help searchers write more precise queries. In an operational environment, the role of a thesaurus can be evaluated with respect to indexing or to retrieval via reformulation of queries. In this report, we describe the results of an experiment designed to measure the effects of automatic query expansion on retrieval effectiveness. This experiment has been designed to establish a framework on which we can base further experiments.

There are two reports in the information retrieval literature of experiments conducted to determine the usefulness of thesauri[11, 10]. The authors describe experiments that were conducted using two different document collections and two different environments. These reports present contradictory results.

In Section 2 of this report, we give a historical perspective by reviewing the experiments reported in the literature. In Section 3 we describe the document and query collection used in our experiment. We then explain the test environment in Section 4. In Section 5 we describe our experiment and present the results. Finally, in Section 6 we give conclusions and suggestions for future work.

2 Historical Perspective

A well designed thesaurus can be very useful in subject searching in an online document collection. It is an irreplaceable tool as an aid in manual indexing. An indexer can consult the thesaurus to assign appropriate terms to individual documents. It is also invaluable for formulating good search strategies. Thesauri can be used to retrieve more relevant documents (i.e. increasing recall) by expanding the query with related terms. It can also be used to suppress retrieval of non-relevant documents (i.e. increasing precision) by using narrower terms. It is this manual indexing and search strategy that has led to an abundance of manually constructed thesauri.

Because of the importance of thesauri for manual indexing and query reformulation, we would like to discover whether the LSS thesaurus can be used in an automated fashion to increase precision and recall. The literature presents two contradictory points of view using very different environments. We briefly describe these experiments and relate them to the one we have performed on the LSS prototype collection.

2.1 Thesauri Experiment at IIT, 1985

This experiment was one of the first done to evaluate the automated use of relational thesauri in information retrieval[11]. It is a fairly small experiment and is typical of many experiments

reported in the field of information retrieval with small collections. Mainly done as a class project at the Illinois Institute of Technology (IIT), the document collection was a small set of abstracts from the *Communications of the ACM*. The queries, thesaurus, and relevancy judgments were all designed by the faculty and students of IIT. There were a total of 29 queries. Ten of these queries were ill-formed (i.e. very few search terms). The results showed that the use of a thesaurus could improve both precision and recall. The experiment also implied that the best improvement came with the ten ill-formed queries. Although the results they report showed statistically significant improvement in precision and recall, the size of the collection and its environment make the results questionable.

2.2 Query Expansion with WordNet, 1994

The field of information retrieval has an extensive list of experiments on small and manageable collections. Many of these experiments suggest improvement in precision and recall that do not yield the same improvement in large collections such as TREC[2]. Since the inception of TREC in 1994, many researchers have been discovering that some of the past techniques need to be retested with the large collections. One such case is the retesting of the above-mentioned experiment[10]. In this new experiment, the queries were expanded using WordNet[4], a manually constructed lexical system at Princeton University. The experimental results showed that this query expansion made little difference in precision and recall. While some queries showed improvement, others showed degradation. On average, the use of WordNet did not improve retrieval effectiveness in the TREC collection.

3 Document and Query Collection

Our collection consists of 674 documents (26,467 pages) which are part of the LSS prototype document collection that was given to the Information Science Research Institute (ISRI) by the Department of Energy (DOE) for continued study in the areas of optical character recognition and information retrieval. The full collection consists of approximately 2,600 documents (104,000 pages) together with their corrected ASCII text and page images. These documents make up the Licensing Support System (LSS) prototype.

As would be expected from this kind of collection, most of the documents deal with technical, scientific topics. Although the documents tend toward the scientific, within this domain the collection is diverse. Our set includes topics from rock mining to safety issues for the transportation of nuclear waste. The set we use covers all sixteen established subject areas designated for the LSS.

This collection consists of full-text documents. The average document length is forty pages and the median length is sixteen pages. For more information about the complete document collection, see [5].

The queries we use for our testing are a subset of the LSS prototype test questions. These queries were artificially constructed to evaluate how well users were able to retrieve needed information from the database. Many of the queries were written to retrieve information from the structured fields of the records, not the actual text. Some of these structured fields are: author name, title, descriptor field, and document type. Because of this difference, some of the original queries were excluded from our test set; many others were reworded so as to reference only the text of the document. The translation of the original English queries to their FQM representation was done by a geologist, two computer scientists, and

Test Query INJD-T3-Q1

LSS Prototype Test Question: Your office is trying to trace the evolution of NRC’s position on repository sealing concepts (e.g. shaft and borehole seals). You need to produce a listing of all documents (including meeting material) discussing seals.

Text only translation: *Find documents discussing repository sealing concepts (shaft and borehole seals).*

FQM translation: find document where text include phrase like 'repository' & 'seal' or text include phrase like 'shaft' & 'seal' or text include phrase like 'borehole' and 'seal' order by docid

Figure 1: Example test query translation

two research assistants to ensure correctness. The interpretation of the original queries was not lost and they represent an unbiased set of sixty-eight queries. Figure 1 is an example of an original test query, its text-only interpretation, and its translation in BASISplus’ query language, FQM.

The relevancy judgments were not part of the collection we received from the DOE. We collected this information for this document set during our OCR and text retrieval research for the Information Science Research Institute[9, 7, 8]. The queries were examined and the complete set of documents was divided among a group of geology graduate students to determine relevance. The students made binary relevancy judgments, classifying documents as either relevant or not relevant. Excluding seven of the queries having no relevant documents in the collection, the average number of relevant documents per query is 10.

4 The BASISplus Environment

4.1 Full Text Retrieval

BASISplus Release L is the text retrieval system we used for our experimentation. BASISplus’ document database uses traditional Boolean logic positional inverted file methodology[1]. The most commonly used method for extracting information from an inverted file text retrieval system is a Boolean logic query language[6]. These languages can be used to locate document information using combinations of Boolean expressions or *queries*. The answers to the queries are computed by manipulating the stored information for the terms¹ that appear in each query. Additional query language features include wildcarding, stemming and canonical forms, proximity searching, and access to thesauri. These features are used to try to improve precision and recall.

BASISplus offers several parameters for document loading. Certain parameters affect retrieval results, in particular the `Search_Control_Set`, which specifies search criteria. The parameters we use in our experimentation are defined below:

- `Break_List` defines which characters delimit words. Each word becomes a term in the document collections’ index. We use `non_graphic` and `non_textual` lists defined

¹We use *term* and *word* interchangeably.

by BASISplus which include all non-printable characters and all printable characters excluding letters, digits, and hyphens, respectively.

- **Sub_Break_List** is used in conjunction with the **Break_List** to aid searching on words joined by special characters such as hyphens. We include only forward slashes (/) and hyphens (-) in this list. The **Sub_Break_List** causes subterms and superterms to be included in the document collection index.
- **Stop_Word_List** identifies common words that are of no value for document retrieval (e.g. the, and, or) and so are not indexed. We use the system-defined **Stop_Word_List** as documented in [3].
- **Raise_Terms** raises both the search terms and the data in the index causing case to become irrelevant. We set this to **YES** to ensure that Yucca Mountain is located even if it appears as YUCCA MOUNTAIN in the document's text.
- **Singular** converts plural words to singular. The method used by BASISplus is not very sophisticated and may actually degrade retrieval effectiveness. We set this parameter to **NO**.
- **Thesaurus** indicates the thesaurus database the system should apply for word association during document indexing. The thesaurus itself has several options which we will cover in more depth in Section 4.3.

4.2 Query Environment

BASISplus provides a query language called *Fundamental Query and Manipulation* or FQM. FQM is a command language based on Boolean logic that supports wildcarding and proximity searching. These features are used infrequently in our queries. One of our queries uses wildcarding and only phrase proximity searching is employed.

Although FQM is based on first order logic, its syntax differs from the SQL standard. We found it impractical and cumbersome for our testing environment and in general, we could see potential difficulties for searchers once the LSS retrieval database was in place. That said, we point out that we are using Version L, the 1990 release of BASISplus. Improvements to this system may have been made since this release.

4.3 LSS Thesaurus

The LSS thesaurus builds relationships between words specific to the Licensing Support System's domain. Most thesauri, including the LSS thesaurus used in this experiment, relate terms using *relation types*. Relation types indicate the type of relationship between two or more terms in the thesaurus. The BASISplus thesaurus module provides the following relation types:

Equivalent, Preferred/Non-preferred Usage. These would include synonyms, abbreviations and specially coded terms. For example, **DOE** and **Department of Energy** are synonymous; both are correct and appear throughout the collection's text. Searchers may search on either term. The thesaurus aids queries that use these terms by setting up a preferred relationship between them. Either term can be searched for and documents containing either term will be returned.

Broader/Narrower, General/Specific Usage. These relationships organize thesaurus term hierarchy. Hierarchical thesauri can retrieve specific topics implied when more general search terms are used. SEDIMENTARY ROCKS would be considered a broader term of CARBONATE ROCKS whereas CALICHE is a narrower term than CARBONATE ROCKS.

Joining, Component/Composite Usage. This relation type represents a combination of two different ideas within a single search word or phrase. A composite term is made up of two or more component terms. Although these are defined for the thesaurus, no relationships of type component/composite are implemented in the application thesaurus.

Related, Reference Usage. Reference usage indicates that some logical relationship exists between two terms. For example, SEDIMENTS is related to ALLUVIAL DEPOSITS but there is no hierarchical relationship between them.

Definition Notes, Informative Usage. This relation type allows those who construct the thesaurus to add informative definitions to aid the user. EROSION for example includes the following definition note in the thesaurus:

The effects of natural agents acting on the land causing destruction and/or removal of material by water, moving ice, or wind.

Building a complete and useful thesaurus for the Licensing Support System requires specialists with a deep understanding of the LSS document collection and of its scientific content. BASISplus allows several ways to apply this thesaurus to a document collection. These will be discussed in Section 4.4.

4.4 Applying Thesaurus Control

In BASISplus, the Database Administrator specifies rules for thesaurus control in the form of a *thesaurus definition* at the time the document collection is loaded. This definition is crucial to the thesaurus' serviceability for those who eventually query the system. Two types of rules can exist in the definition:

Rules that control data entry. Data entry rules actually replace nonpreferred terms in the document collection with preferred terms as indicated in the thesaurus. As previously mentioned in Section 4.3, preferred terms are terms that are synonymous with nonpreferred terms.

Rules that control retrieval. Retrieval control rules can replace nonpreferred query search terms with preferred search terms or can combine or expand a query with selected thesaurus terms. The selected thesaurus terms can include terms from any of the relation types listed in Section 4.3.

BASISplus allows any combination of rule types but warns that only certain combinations will give desired results. For this preliminary study, we use *Combined Retrieval Control* to expand the queries with preferred terms only.

Query	Thesaurus	Documents Returned	Relevant Returned	Recall	Precision
INJC-T1-Q2	Excluded	39	5	0.83	0.13
	Included	41	5	0.83	0.12
INJC-T3-Q2	Excluded	22	4	0.12	0.18
	Included	30	7	0.21	0.23
INJD-T1-Q1	Excluded	10	1	0.12	0.10
	Included	17	1	0.12	0.06
PIJA-T2-Q1	Excluded	15	0	0.00	0.00
	Included	16	0	0.00	0.00
PIJC-T3-Q3	Excluded	4	0	0.00	0.00
	Included	7	0	0.00	0.00
PIJD-T3-Q2	Excluded	24	3	0.11	0.12
	Included	96	6	0.22	0.06
RLJA-T1-Q2	Excluded	25	8	0.17	0.32
	Included	90	27	0.57	0.30
TEJA-T3-Q1	Excluded	66	6	0.33	0.09
	Included	71	7	0.39	0.10
TEJD-T1-Q1	Excluded	3	0	0.00	0.00
	Included	4	0	0.00	0.00

Table 1: Nine queries with increased recall using the LSS thesaurus

5 Preliminary Study

With a collection as voluminous and comprehensive as the LSS, the Department of Energy should justifiably be concerned with retrieval effectiveness. The preliminary study we present gives a baseline for investigating ways to improve the LSS database’s performance and for providing guidance for further study within their specific domain.

Two identical document sets were loaded into BASISplus. Both sets used the same parameters for document loading with the exception of thesaurus control. In one set we applied thesaurus control, in the other we did not. We will refer to these sets as *thesaurus-included* and *thesaurus-excluded* respectively.

The same set of 68 queries were batch run against each database. Fifty-nine queries return exactly the same results. The remaining nine queries in the thesaurus-included set showed an increase in recall. The total number of documents returned for these queries, the number of relevant documents returned, and their recall and precision are shown in Table 1.

As pointed out in Section 4.4, we only applied Combined Retrieval Control to expand the queries with preferred terms. With this type of control, recall should be increased. Note that the results for the queries run against the thesaurus-included set returned either the identical results as thesaurus-excluded or returned a superset of documents when compared to the results in thesaurus-excluded. Combined Retrieval Control does not reduce the number of non-relevant documents.

The nine queries returning a superset of documents contained search terms that were considered *nonpreferred* in the LSS thesaurus (Table 1). These are the only queries that

Thesaurus	Recall	Precision
Excluded	0.34	0.10
Included	0.35	0.10

Table 2: Average precision and recall for the complete set of queries

could be improved with the parameter selection we applied.

Average precision and recall for the complete set shows no significant difference. The results are shown in Table 2. Average results tend to mask individual improvements; in some individual queries, recall was affected dramatically. For example,

RLJA-T1-Q2 returned 27 relevant documents in the thesaurus-included set compared to 8 in thesaurus-excluded, increasing recall from 17% to 57%.

INJC-T3-Q2 nearly doubled the number of relevant documents returned from 4 in thesaurus-excluded to 7 in thesaurus-included.

With thesaurus-aided retrieval, the above queries were able to retrieve several more relevant documents. As discussed in [11] these queries may be considered *ill-formed*. They were short, with fewer than five search terms each. But this makes sense; if a query is initially written complete with synonyms and combined with choice terms, a thesaurus would be of little value in an automated environment. What these results show is that there is potential to improve query effectiveness when thesauri are applied to the document collection. And although only nine queries show increased recall, this represents almost 15% of the queries run. Expanding thesauri testing will assist in discovering its potential.

We did attempt to run the experiment with various combinations of the rules available in BASISplus. Unfortunately, this version of BASISplus was unable to accommodate some query expansions. This was the case even when these queries were run interactively in FQM. BASISplus is a large complex system that has a variety of extensions and add-on features. While these additions are sometimes useful, they may cause complications and frustration for many users. The consequence of search control design by the Database Administrator cannot be overstated.

6 Conclusion and Future Research

This experiment indicates that combining one thesaurus relation type, namely preferred terms, has no significant effect on average precision and recall. There is no question that other relation types should be examined as well—alone and in combination. In particular, it would be interesting to know if there is some combination of relations that can improve retrieval effectiveness for the LSS collection within the designated environment.

The result of this experiment is inherently dependent on the environment we employed. There are many aspects of this environment that can affect our findings. The two cited experiments in the literature[11, 10] deal with two different collections. The noticeable differences in these collections are their size and specificity. The experiment at the Illinois Institute of Technology deals with a small specific collection, while the TREC experiment is based on a much larger collection using a more diverse set of documents. The LSS collection that we use is very specific but small in comparison to the TREC collection. Future

experiments should expand the experimental collection size to a more realistic reflection of the eventual LSS environment.

Another aspect of the collection we should take into consideration is that many of the LSS documents will be generated using OCR. An interesting question arises with respect to thesauri applications. “Would controlled indexing offered by thesauri compensate for misrecognized words in OCR text?” There are several idiosyncrasies of OCR text that may be improved by applying thesauri. In several of our previous experiments[9, 7, 8], we have shown that if OCR text can be filtered, improvements can be gained in retrieval effectiveness. A thesaurus could aid in this filtering process.

The use of BASISplus as our retrieval system implies that our results hold for operational environments based on the Boolean logic model. It remains to be seen whether these results also hold true with respect to statistically based models such as vector space and probabilistic. Comparing the appropriateness of thesauri use for the LSS collection within these other operational environments would also be worthy of study.

Finally, this preliminary study judges the effect of the LSS thesaurus on retrieval in an environment with no user input. Can an experienced searcher use his/her knowledge of the collection and the richness of a thesaurus to improve results? After all, writing a good Boolean query is not an easy task. A conceptually built thesaurus may aid an experienced searcher in reformulating his/her thoughts when writing queries. Future experimentation could also address this issue.

Although our experiment has shown no improvement in retrieval effectiveness from the use of a thesaurus, the authors believe that the use of a thesaurus can still be of benefit to document searchers.

References

- [1] Delphi Consulting Group, Inc. Text retrieval systems: A market and technology assessment, 1991.
- [2] D. K. Harman, editor. *The Text REtrieval Conferences (TREC 1-5)*, Gaithersburg, MD, 1992-1996. National Institute of Standards and Technology.
- [3] Information Dimensions, Inc. *The BASISplus Retrieval System Manual*. Dublin, OH, 1990.
- [4] George Miller. Special issue, WordNet: An on-line lexical database. *Intl. Journal of Lexicography*, 3(4), 1990.
- [5] T. A. Nartker, R. B. Bradford, and B. A. Cerny. A preliminary report on UNLV/GT1: A database for ground-truth testing in document analysis and character recognition. In *Proc. 1st Symp. on Document Analysis and Information Retrieval*, pages 300–315, Las Vegas, NV, March 1992.
- [6] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw Hill, New York, 1983.
- [7] Kazem Taghva, Julie Borsack, and Allen Condit. Results of applying probabilistic IR to OCR text. In *Proc. 17th Intl. ACM/SIGIR Conf. on Research and Development in Information Retrieval*, pages 202–211, Dublin, Ireland, July 1994.

- [8] Kazem Taghva, Julie Borsack, and Allen Condit. Effects of OCR errors on ranking and feedback using the vector space model. *Inf. Proc. and Management*, 32(3):317–327, 1996.
- [9] Kazem Taghva, Julie Borsack, Allen Condit, and Srinivas Erva. The effects of noisy data on text retrieval. *J. American Soc. for Inf. Sci.*, 45(1):50–58, January 1994.
- [10] Ellen M. Voorhees. On expanding query vectors with lexically related words. In D. K. Harman, editor, *Proc. of the Second Text REtrieval Conference (TREC-2)*, pages 223–231, Gaithersburg, MD, March 1994. National Institute of Standards and Technology.
- [11] Yih-Chen Wang, James Vandendorpe, and Martha Evens. Relational thesauri in information retrieval. *J. American Soc. for Inf. Sci.*, 36(1):15–27, 1985.