An Association Thesaurus for Information Retrieval

Yufeng Jing and W. Bruce Croft
Department of Computer Science,
University of Massachusetts at Amherst,
Amherst, MA 01003.
jing@cs.umass.edu, croft@cs.umass.edu

Abstract

Although commonly used in both commercial and experimental information retrieval systems, thesauri have not demonstrated consistent benefits for retrieval performance, and it is difficult to construct a thesaurus automatically for large text databases. In this paper, an approach, called PhraseFinder, is proposed to construct collection-dependent association thesauri automatically using large full-text document collections. The association thesaurus can be accessed through natural language queries in INQUERY, an information retrieval system based on the probabilistic inference network. Experiments are conducted in INQUERY to evaluate different types of association thesauri, and thesauri constructed for a variety of collections.

1 Introduction

A thesaurus is a set of items (phrases or words) plus a set of relations between these items. Although thesauri are commonly used in both commercial and experimental IR systems, experiments have shown inconsistent effects on retrieval effectiveness, and there is a lack of viable approaches for constructing a thesaurus automatically. There are three basic issues related to thesauri in IR as follows. These issues should each be addressed separately in order to improve retrieval effectiveness.

- **Construction:** There are two types of thesauri, manual and automatic. The focus of this paper is on how to construct a thesaurus automatically.
- **Access:** Given a particular query, the thesaurus must be accessed and used in some way to improve or expand the query.
- **Evaluation:** After a thesaurus is built, it is important to know how good it is. Manual thesauri are evaluated in terms of the soundness, coverage of classification and thesaurus item selection. The evaluation of automatic thesauri is generally done via query expansion to see if retrieval performance is improved.

There are two types of manual thesauri. The first are general-purpose and word-based thesauri like Roget's and WordNet. Those thesauri contain sense relations like antonym and synonym but are rarely used in IR systems. The second are IR-oriented and phrase-based thesauri like INSPEC, LCSH (Library of Congress Subject Headings), and MeSH (Medical Subject Headings). Those manual thesauri usually contain relations between thesaurus items such as BT (Broader Term), NT (Narrow Term), UF (Used For), and RT (Related To), and can be either general or specific, depending on the needs of thesaurus builders. This type of manual thesauri is widely used in commercial systems. The major problem with manual thesauri is that they are expensive to build and hard to update in a timely manner. Even though the determination of thesaurus item relations is made by human experts, it is a difficult task. For example, what is the relation between "information system" and "data base system"? Although generally built to support information retrieval, manual thesauri have been evaluated by testing the soundness and coverage of the thesaurus concept classification. This does not always directly serve the purpose of information retrieval. Good classification and concept coverage do not guarantee effective retrieval.

An automatic thesaurus is usually collection-dependent, i.e. dependent on the text database which is used. A few small and automatically constructed thesauri have been used in experimental IR systems but the effectiveness of these thesauri has not been established for large text databases. Automatic thesauri are typically built based on co-occurrence information, and relevance judgements are often used to estimate the probability that thesaurus terms are similar to query terms or a particular query. Since determining term or phrase relations is hard to achieve automatically, in automatic thesauri these relations are simplified to one type of association relation.

In this paper, an approach for the automatic construction of thesauri is presented along with a program, PhraseFinder, that utilizes this approach to construct a collection-dependent thesaurus, called an **association thesaurus**. The association thesaurus is accessed through INQUERY, an information retrieval system based on inference networks, using natural language queries. Adding phrases retrieved in this way to the original queries produces significant performance improvements. In the following sections, previous work is reviewed, the approach and techniques for PhraseFinder are described, experimental results are presented, and some future research issues related to association thesauri are discussed.

2 Thesauri and Information Retrieval

The use of thesauri in IR involves automatic construction, user interface design, retrieval mechanisms, and retrieval architecture. Research related to automatic thesauri dates back to Spark-Jones's work on automatic term classification [18, 19], Salton's work on automatic thesaurus construction and query expansion [14, 16], and Van Rijsbergen's work on term co-occurrence [10]. In those experiments, automatic term classification without relevance judgments or feedback information did not produce any significant improvements. Minker, Wilson, and Zimmerman in [9] after an exhaustive investigation came to the same conclusion.

Salton [14] showed that using the Harris synonym thesaurus with relevance judgements produces significant improvements. He proposed two approaches : term-document and term-property matrices, for automatic construction of thesauri based on relevance judgements. In [16, 15, 22] Salton, Yu, and Buckley further developed these methods into a formal term dependence model. But the serious drawback with the term dependence model is that it assumes the availability of relevance judgements. The occurrence information of terms in relevant and non-relevant documents is used to estimate the probability of terms similar to queries. When full relevance judgments are not fully available, the retrieval performance under the model degraded badly [22]. Recently Crouch and Yang[4, 5] used Salton's approach to build a term-vs-term thesaurus whose classes are clustered in terms of discrimination values. Although such a thesaurus improved retrieval performance, to determine the best threshold value is difficult because such a value is determined assuming the availability of relevance judgements. Yu's study on clustering for information retrieval [23] showed that term dependence information improves the effectiveness of retrieval systems. In the non-binary independence model, he used hierarchical term relations.

Rijsbergen and Smeaton in [10, 17] used a statistically-derived MST (*maximum spanning tree*) as a term dependence structure with relevance feedback to verify the following *association hypothesis*

> *If an index term is good at discriminating relevant from non-relevant documents then any closely associated index term is also likely to be good at this.*

The experiments using MST resulted in no positive results.

In the recent work by Yonggang Qiu and H.P. Frei[13], a term-vs-term similarity matrix was constructed based on how the terms of the collection are indexed. A probabilistic method is used to estimate the probability of a term similar to a given query in the vector space model. Even when adding hundreds of terms into queries, this approach showed that the similarity thesaurus can improve performance significantly. Adding so many terms may not, however, be efficient for large information retrieval systems. It should be noted that their improved results on the NPL collection is still lower than the baselines used in this paper.

In the CLARIT project[6], NLP techniques are used to identify *candidate noun phrases* in full-text and map them into *candidate terms*, in a morphologically-normalized form, emphasizing *modifier* and *head* relations, and the candidate terms are matched against a *first-order thesaurus* of certified domain-specific terminology, which is a semiautomatically generated control vocabulary dictionary. This indexing technique builds a mapping between terms within a candidate noun phrase and terms in the first-order thesaurus. As an indexing technique, the method used in CLARIT was compared with human indexers on the basis of ten documents. Ruge[11] uses the *head/modifier relation* to investigate the effect of term associations on the performance of IR systems in the hyperterm system REALIST. But this type of association might be too restrictive to capture some associations, for example, for the query "insider trading scandal", person and company names which were involved in and other phrases which often co-occur with "insider trading scandal" can be very important associated phrases.

In summary, some problems with previous work are

- Relevance judgements are used to estimate the probability of a term related to another term or query. Because relevance judgements are often not available, these approaches are impractical for term classification or thesaurus construction. Second, even if available, relevance judgments are often produced for a set of queries, which do not cover a whole collection.

- Term selection for query expansion is based on individual terms but not on the whole query. This is consistent with the association hypothesis, but it ignores the additional context that additional terms in the query provide. A term may be closely associated with one of the words in the query, but not with the meaning of the entire query.

- Co-occurrence data for terms is gathered from individual texts without limiting the range of the co-occurrence. Previous research has focussed on short, abstract-length documents. With full-text databases, where documents may be many pages in length, it is important to consider "text windows" that restrict the co-occurrences.

- All words (verbs, adjectives, adverbs, and nouns) are treated equally, i.e. an implicit assumption is that that all words contribute equivalently to retrieval performance.

The following section describes an approach for automatic thesaurus construction that does not use relevance judgments or relevance feedback. Instead of term clustering or classification, a retrieval system is used to measure the probability of thesaurus terms being associated with a particular query. The top-ranked terms are used to automatically expand the query.

3 PhraseFinder

PhraseFinder is a program that automatically constructs a thesaurus using text analysis and text feature recognition. It is designed to gather data about associations between phrases and terms over a large amount of text, and views a text as a structured object. Text objects consist of paragraphs which consist of a series of sentences which contains phrases and words. Terms are defined as all words except stop words. Within a text, terms are basic features, and the composite features are paragraphs, sentences, and phrases, called *"characteristic features"* of the text. Noun phrases are used as the main characteristic features instead of terms because there is evidence that they play an important role in characterizing the content of a text [3, 1].

PhraseFinder considers co-occurrences between phrases and terms as associations. As we will see later, phrases do not have to be noun phrases. They are defined by a set of rules. In an association thesaurus, the association between phrases and terms is extracted and represented as a text feature dependence. The following basic text features are used in PhraseFinder:

- **Terms** - Any word except a stop word is a term.
- **Part of Speech** - Each word has a part of speech that was assigned by a part-of-speech tagger. So far, the Church tagger[2] has been used to tag texts.

In addition, PhraseFinder recognizes phrases based on a set of simple phrase rules and the part of speech of terms. The following text composite features are used :

- **Paragraphs** - A paragraph consists of a set of sentences. It can be a natural text paragraph or a fixed number of sentences.
- **Sentences** - A sentence is a sequence of words, whose end is recognized by the tagger.
- **Phrases** - A phrase is a sequence of terms whose part-of-speech satisfy one of the specified *phrase rules.* For example, the simple noun phrase rule {NNN, NN, N} means that a phrase can be triple nouns, double nouns, or single noun. PhraseFinder does not allow a phrase to cross sentence boundaries.

PhraseFinder, as its name implies, identifies the associations between phrases and terms so that associated phrases can be retrieved through natural language queries and used for expansion. Regular stemming (i.e. the Porter stemmer[12]) is used for terms. Because this stemmer is quite "aggressive" in terms of removing endings and does not produce real word forms as output, we used a more conservative approach for stemming phrases [8].

3.1 Association Generation and Paragraph Limits

Associations represent the co-occurrence between terms and phrases within texts. Before the generation of associations is described, an important question is what range should be used to generate associations. Intuitively, a natural paragraph seems appropriate because it generally focuses on describing an individual topic. Because very long paragraphs would generate large numbers of associations that would have less chance of being valid, we restricted the maximum number of sentences allowed within a paragraph (known as the *paragraph limit*). Later experimental results show that 3-10 sentences per paragraph is sufficient for full texts.

The procedure for association generation is quite simple. Sentence by sentence, PhraseFinder reads texts, recognizes phrases and terms, inserts phrases and terms into hash dictionaries, stores phrase and term identifiers separately into tables along with their occurrence frequencies, until the number of read-in sentences reaches the paragraph limit or the end of a paragraph is reached. After all phrases and terms have been obtained within a paragraph, pairwise associations are generated between terms and phrases. The association frequency, which is equal to term frequency times phrase frequency, is also stored with each association. An association is a triple:

$$< termid, phraseid, association\_frequency >.$$

PhraseFinder goes through all texts in a collection and generates association data. The association frequency for an association is summed over a whole collection.

3.2 Statistics and Association Data Filtering

The amount of association data is very large. Many associations are, however, infrequent. This suggests that data filtering is necessary. Some of the information that can be used for filtering includes association frequency between phrases and terms, the number of associated terms or phrases for a phrase or term, the total of association frequencies of associated terms or phrases for a phrase or term, the document and collection frequencies of terms or phrases. Some simple statistics based on the association data for simple noun phrases are:

- Around 70% of association data has the association frequency of 1 over TIPSTER volume 1.

- Around 90% of phrases occur just once in only one document.

In addition, a small number of phrases (especially single nouns in our experiments) occur in many documents. These phrases are associated with many terms. Phrases such as *people, company*, and *state,* are too general and not useful for identifying a topic.

On the basis of these observations, we do data filtering on association data and produce an acceptable association thesaurus by discarding associations with frequency 1 and phrases that are associated with too many terms. The latter parameter is set experimentally for each collection. As it turns out, data filtering not only reduces the amount of association data but also enhances retrieval performance.

3.3 Access to an Association Thesaurus

As mentioned early, thesaurus access is handled separately from thesaurus construction. Thesaurus access is a procedure that measures the closeness of thesaurus items in the context of a particular query. In this paper, we implement PhraseFinder access through INQUERY, an information retrieval system based on the probabilistic inference network [21]. To do this, an association thesaurus is converted into a form suitable for INQUERY. Specifically, thesaurus phrases are mapped into pseudo-INQUERY documents where the representation of a pseudo-document consists of all the associated terms. The PhraseFinder version of INQUERY can then accept natural language queries and output a ranked list of phrases (rather than documents) associated with the query. This list is used for automatic query expansion. Figure 1 shows an example query from the TIPSTER collection and the top 20 phrases that are retrieved by PhraseFinder. The ranking values are not used as the weights of the associated thesaurus items.

```
Query:115.1 : Impact of the 1986 Immigration Law  - will report
specific consequence consequences of the U.S.'s Immigration Reform
and Control Act of 1986.

   0.511462        illegal immigration
   0.501936        illegals
   0.499120        undocumented aliens
   0.498964        amnesty program
   0.498054        immigration reform law
   0.492453        editorial-page article
   0.490993        naturalization service
   0.489448        civil fines
   0.488754        new immigration law
   0.487762        legal immigration
   0.487187        employer sanctions
   0.483245        simpson-mazzoli immigration reform
   0.482687        statutes
   0.480449        applicability
   0.480222        seeking amnesty
   0.478625        legal status
   0.478437        immigration act
   0.477798        undocumented workers
   0.475995        guest worker
   0.475995        sweeping immigration law
```

Figure 1: Example PhraseFinder Output

4 Query Expansion Using An Association Thesaurus

The evaluation of a thesaurus is an open issue. There are some criteria to evaluate the quality and category soundness of a manual thesaurus, but these criteria do not predict whether the manual thesaurus will improve retrieval effectiveness.

Automatic thesauri have been evaluated by doing query expansion and measuring retrieval effectiveness. In query expansion, thesaurus items are retrieved for a particular query and are added into the query. To some extent, query expansion shows whether or not a thesaurus can provide useful phrases (or terms for term-based thesauri) for searchers. To measure the performance of PhraseFinder, we conducted experiments using thesauri built from the NPL and TIPSTER collections [7]:

- NPL: Associations generated from 11,429 documents containing titles and/or abstracts in the area of physics.
- TIPSAMP: Associations generated from a sample of the TIPSTER collection consisting of about 52,000 full-text documents from AP, ZIFF, and WSJ (87, 88, 89). It was produced by taking every fifth document out of TIPSTER volume 1. It contained about 15,000

AP documents, 17,000 ZIFF articles, 20,000 WSJ articles. Federal Register and DOE are part of TIPSTER volume 1 but were not used in this version of PhraseFinder because very few queries are directed at these documents.[1]

- TIPFULL: Generated from about 250,000 full text TIPSTER documents containing about 85,000 AP news, 75,000 ZIFF articles, and 93,000 WSJ (87, 88, 89) journal articles.

Obviously, these collections vary dramatically from text length to content. All documents are tagged, and PhraseFinder takes tagged documents as input to produce association data. It takes two weeks to generate an association thesaurus for TIPFULL, a week for TIPSAMP, and hours for NPL.

In these experiments, query expansion was done by using a natural language query (or part of the query in the case of TIPSTER) to retrieve a ranked list of phrases from the PhraseFinder version of INQUERY. Given this ranked list, the following methods were used to determine which phrases to add to the original query.

- **Only Duplicates**

  Only duplicate phrases from an association thesaurus are added into queries. A phrase is a duplicate for a given query if all terms (simple stemming is used) of the phrase are subsets of the query. E.g. for query "*high frequency oscillators using transistors*", phrases "*transistor oscillator*" and "*frequency oscillator*", and terms "*transistors*" and "*frequencies*" all are duplicate with the query. The purpose of adding duplicate phrases is to see whether the evidence from the thesaurus identifies the importance of query terms or phrases.

- **Nonduplicates**

  Nonduplicate phrases are added into queries. If any term of a phrase is not a subset of a given query, the phrase is nonduplicate with the query. For the above example query, phrases "*npn transistors*" and "*junction transistor*", and terms "*triode*" and "*anode*" are nonduplicate with the query. The purpose of adding nonduplicate phrases is to see whether a thesaurus can provide some new information about queries.

- **Both Duplicates and Nonduplicates**

  Both duplicate and nonduplicate phrases are added into queries.

Both weighting and non-weighting schemes are used for query expansion. The following conventions are used to indicate how queries are expanded:

BOTH	adding both duplicate and nonduplicate phrases
DUPONLY	Only adding duplicate phrases
NODUP	Adding nonduplicate phrases

The top 50 ranked associated phrases are used as candidates for query expansion. In runs that used duplicate phrases, all duplicates in the top 50 were used. The number of nonduplicate phrases and the weighting of duplicate and nonduplicate phrases were varied in a large number of experiments. The results reported here are summaries of those experiments and only a few examples of these parameters are given.

[1] *WSJ* stands for **Wall Street Journal** articles, *AP* for **Associated Press Newswire**, *ZIFF* for **ZIFF articles**, *doe* for short abstracts from **Department of Energy**.

5 Experiments on Small Collections

The phrase rule in PhraseFinder does not have to define linguistic phrases. In fact, it defines what type of "concept" associates with terms. Different phrase rules result in different types of association thesauri produced by PhraseFinder. Terms are classified into several classes, each of which are symbolized by individual letter tokens as follows, in terms of part of speech.

token	meaning
N	noun
J	adjective
R	adverb
V	verb
G	verb-ing
D	verb-ed
I	cardinal number

The purpose of this classification is to investigate retrieval effectiveness on different types of words for query expansion. A phrase rule is a set of sequences of above tokens. PhraseFinder produces different types of association thesauri for different phrase rules. Experiments were conducted to address the following issues:

- How do different types of words, like verbs, adjectives, adverbs, and nouns affect retrieval performance? This question may indirectly relate to the role of different types of words in conveying the content of texts.
- What are the best phrases for an information retrieval system like INQUERY?
- Comparison of a word-based thesaurus and a noun-phrase-based thesaurus.

In following subsections, two types of association thesauri: word-based and noun-phrase-based association thesauri are constructed on the NPL collection, evaluated, and compared.

5.1 Word-based Association Thesauri

A word-based association thesaurus means that association data generated by PhraseFinder is between different type of words and terms. Most automatic thesauri are based on term-vs-term co-occurrence, called a term-based thesaurus here. PhraseFinder is able to produce different kinds of thesauri based on its phrase rule. For example, to produce a term-based association thesaurus, the phrase rule can be defined as {N, J, R, V, G, D, I}, i.e. all terms are "phrases". Words are categorized into three classes: verbs, adjectives and adverb, and nouns. Thesauri for different classes of words are generated on the NPL, and these thesauri are evaluated and compared.

Intuitively, if all nouns are removed from a text, it is generally hard to understand the text. If all verbs are removed from a text, it is often possible to know what the text is about. A word-based thesaurus can be a way to test how words with different part of speech affects retrieval performance and how good they are at representing the content of texts. Obviously these experimental results do not exclude the possibility that some words are particularly important in conveying the content of some texts. The term-based thesaurus, which ignores part of speech, is also evaluated and compared with the others. The brief summary of experimental results on word-based association thesauri is reported in Table 1, in which only improvement percentages over a common baseline are presented. These results represent the best performance over a range of parameter settings, such as weighting of added phrases and number of added phrases.

Thesaurus Type	Phrase Rule	Methods for Query Expansion (%)		
		DUPONLY	**NODUP**	**BOTH**
Verb-based	{V,G,D}	+0.1	+0.8	-0.2
Adj+Adv-based	{J, R}	+2.2	+1.7	+3.0
Noun-based	{ N }	+4.0	+0.9	+4.2
Term-based	{N,J,R,V,G,D,I}	+1.8	+2.6	+2.4

Table 1: Brief Summary of Results on Word-Based Association Thesauri on NPL

When phrase rule is {V, G, D}, a verb-based association thesaurus is generated. The associations between verbs and terms are generated for the verb-based thesaurus. The results show that verbs do improve retrieval performance a little but not significantly. Reweighting verbs in queries does not make a difference. Adding associated verbs, not already in queries, helps more than reweighting, but 0.8% improvement is not significant either. The result for method 3 (**BOTH**) results in no improvement.

For the thesaurus based on adjectives and adverbs, the phrase rule {J, R} is used to generate association data between adjectives or adverbs and terms. The results show that this thesaurus produces bigger improvements than the verb-based one. That means that reweighting adjectives and adverbs within queries and adding new ones from this thesaurus are mutually complementary.

As expected, the results show that the noun-based thesaurus performs better than either the verb-based one or the one based on adjectives and adverbs over all. Reweighting nouns within queries and adding new nouns associated with queries from this thesaurus are complementary, although it is interesting to see how much of the improvement is due to reweighting or simple variations of existing query terms.

Most existing automatic thesauri are term-based. Comparing with the previous three types of thesaurus, the term-based thesaurus performs less effectively than either the noun-based one or the one based on adjectives and adverbs in INQUERY.

In summary, all types of words are useful for improving retrieval performance but nouns contribute the most, adjectives and adverbs less, and verbs the least. The term-based thesaurus, which ignores part of speech, is even less effective than the one based on adjectives and adverbs.

5.2 Noun-Phrase-based Thesauri

In the following experiments, three sets of phrase rules, corresponding to different phrase forms, are used to produce different noun-phrase-based association thesauri. Table 2 presents the experimental results for these thesauri.

For the thesaurus based on the phrases containing only nouns, whose phrase rule is {NNN, NN, N}, the results in table 2 show that 6.9%, 2.6%, and 9.7% improvements are obtained for three query expansion methods respectively. These improvements are bigger than those on other phrase-based thesauri.

Adjectives can be used to improve performance and many noun phrases are of form adjective + noun phrase. For this reason, a set of more general noun phrase rules, {NNN, JNN, JJN, NN, JN, N}, would be used to generate the association thesaurus. The results in Table 2 show that under the general noun phrase rule, adding duplicate phrases with weight 0.5 produces a 5.5% improvement, adding fifteen nonduplicate phrases with weight 0.1 produces a 3.4% improvement overall. Combining two methods results in a 6.2% improvement. It is clear that adjectives and nouns together do not work as well as nouns alone.

Phrase Rule	Methods for Query Expansion (%)		
	DUPONLY	NODUP	BOTH
{NNN,JNN,JJN,NN,JN,N}	+5.5	+3.4	+6.2
{NNN,NN,N}	+6.9	+2.6	+9.7
{NNN,NN}	+3.5	+2.2	+6.2

Table 2: Summary of Results on Phrase-Based Association Thesauri on NPL

The purpose of removing single noun from phrase rules is to reduce the association data because many phrases are composed of single nouns. It is expected that some important single-noun phrases would be lost because of this. But how big would the loss be? The results with phrase rule {NNN, NN} in table 2, in which duplicate phrases with weight 0.5 and eight nonduplicate phrases with 0.1 are added, show that the loss is significant. It is concluded that single nouns play a very important role in queries and texts. For those single nouns that are very general and not useful for conveying content of texts, simple data filtering techniques can be used to throw them away.

For phrase-based thesauri, the association thesaurus based on noun-only noun phrases produces the best improvement out of all three in INQUERY, and the improvement is significant. The addition of adjectives into noun phrases hurts rather than enhances retrieval performance. Individual nouns are very important noun phrases and help to improve retrieval performance. The experiments show that the noun-phrase-based thesauri perform much better than any type of word-based thesauri.

6 Experiments on the TIPSTER collections

There are three TIPSTER databases built for INQUERY: tip1, tip, and tip3. Table 3 illustrates what is in each database [2].

Database Name	# of docs	Database Content
tip1	510,887	wsj87, wsj88, wsj89, ziff, ap, doe, fr
tip	742,358	wsj87, wsj88, wsj89, ziff, ap, doe, fr, wsj90, wsj91, ziff2, ap2, fr2
tip3	238,848	ziff3a, ziff3b, ap3, patn, sjm_a, sjm_b

Table 3: TIPSTER Database Contents

An original TIPSTER topic (or query) consists of a number of fields, such as the "description", a short natural-language description of the information need, and the "concepts", a list of words and phrases related to the query. Only some of these fields are used in these experiments.

Two sets of TIPSTER topics are used for experiments, which are topics 51-100 and 101-150 respectively notated as **set2** and **set4**. Relevance judgments for the TIPSTER queries are from the TREC-2 conference. For each set of topics, baseline query sets are generated by automatic processing. The query sets used were as follows:

[2] In this table, *fr* stands for **Federal Register**, *patn* for **Patent** data, and *sjm* for the **San Jose Mercury News**. The other abbreviations have been used previously.

- **SET2** There are two sets of baseline queries: set2.qry1 and INQ026. The first is the initial baseline query used in early TIPSTER experiments. The second is the one that so far produces the best results in our TIPSTER experiments. Both set2.qry1 and INQ026 are constructed using all fields of the TIPSTER original queries.

- **SET4** One baseline query set, INQ013, is used. INQ013 is structured using only the description part of the original queries with the phrase operators. The purpose of using INQ013 is to simulate an on-line retrieval environment, in which queries are relatively short.

All thesauri used in these experiments were generated using the phrase rule {NNN, NN, N}. Experiments were conducted to investigate the following issues:

- What the effect of the paragraph limit is on the performance of an association thesaurus and how big this effect would be?

- Is it possible to use a representative sample of a collection to generate a association thesaurus for a whole collection?

- Is it possible to apply an association thesaurus built for one collection to another that overlaps in content?

- How well would an association thesaurus perform in an on-line environment?

6.1 Experiments on the Paragraph Limit

The way that PhraseFinder works, a small paragraph limit can significantly reduce the amount of association data and allow to use fewer computer resources. Experiments for the paragraph limit are to find an appropriate paragraph limit. Since current available standard collections, like NPL and CACM, are not full text, and contain only titles and abstracts, the paragraph limit does not make sense. These experiments with different paragraph limits are conducted on TIPSTER sample collection. As a general observation, if the paragraph limit is too small, e.g. one sentence per paragraph, important associations would be lost. If the paragraph limit is too large, e.g. 15 or 20 or larger, too many associations are generated. A very large paragraph limit means that associations are generated within a natural paragraph.

The concepts field within the original TIPSTER topics were used to retrieve phrases from TIPSAMP thesauri constructed with different paragraph limits, i.e. 3, 5, and 10. Five nonduplicate phrases with weight 0.5 and all duplicate phrases in the list of the top 50 associated phrases with weight 1.0 are added into the queries. Table 4, in which **PL-n** means a paragraph limit of **n**, summarizes the results.

There are some slight changes in performance when the paragraph limit varies from 3 to 10. Using 10 as paragraph limit is a little better than using 3 and 5, using 5 better than 3. This may confirm the intuition that a natural paragraph would be good. However, the differences are not significant. It seems that when a paragraph is limited to 3-10 sentences, the paragraph limit has a little impact on the performance. Even with the high baseline query (i.e. INQ026), a more than 2% improvement is obtained on tip1 and tip, and a 6% improvement on tip3. The improvement on tip3 was unexpected, since it is a different set of documents, from a different time period, than the collection used to build the association thesauri. A related result appears in the next section.

Baseline Queries	TIPSTER Database	baseline query precision (%)	The Paragraph Limit (%)		
			PL-3	PL-5	PL-10
set2.qry1	tip1	37.2	+6.4	+6.6	+7.1
	tip	34.9	+5.4	+5.6	+7.2
	tip3	31.2	+7.4	+8.1	+9.9
INQ026	tip1	41.0	+2.6	+2.5	+2.6
	tip	38.0	+2.6	+2.3	+2.8
	tip3	33.5	+4.7	+4.4	+6.2

Table 4: Summary of the Results on the Paragraph Limit

6.2 Using a Sample Collection

It is currently very inefficient to generate an association thesaurus on all texts in a very large collection. The experiments conducted here test the difference between a thesaurus built using a representative sample of a collection and one built using the whole collection. In the preceding subsection, the TIPSAMP thesaurus for the TIPSTER sample collection was evaluated. Here, the TIPFULL thesaurus with paragraph limit 5 is also used.

	Precision (% change) – 50 queries				
Recall	set2.qry1	TIPSAMP		TIPFULL	
0	83.8	87.2	(+4.0)	86.1	(+2.7)
10	60.5	63.8	(+5.4)	63.9	(+5.6)
20	52.6	54.2	(+2.9)	54.2	(+3.0)
30	46.7	49.6	(+6.3)	49.6	(+6.4)
40	40.5	44.4	(+9.5)	44.5	(+9.7)
50	34.9	39.1	(+12.1)	38.9	(+11.6)
60	30.5	33.6	(+10.4)	32.9	(+8.1)
70	25.3	27.5	(+8.6)	27.5	(+8.7)
80	19.7	20.9	(+6.1)	20.8	(+5.4)
90	12.1	13.3	(+9.4)	13.3	(+9.9)
100	2.4	2.4	(−2.8)	2.4	(+0.1)
average	37.2	39.6	(+6.6)	39.5	(+6.1)

Table 5: Performance of the TIPSAMP and TIPFULL thesauri on tip1

The results in table 5 show that roughly the same improvements are obtained with both thesauri. Even though the thesaurus built using the whole collection certainly contains more associations between phrases and terms, it is clear that in terms of retrieval performance, the association thesaurus for a representative sample of a collection is as good as the one for the whole collection. We have not yet determined, however, what the optimal sample size should be.

In another experiment, we obtained a 10.2% average improvement in precision when the queries expanded using the TIPFULL version of PhraseFinder were evaluated with the tip3 database. This leads to the conclusion that for two collections, an association thesaurus built for one collection can be used for another if the content of the two collections overlaps significantly.

6.3 The Use of An Association Thesaurus in An On-line Environment

In the preceding two subsections, highly-structured queries are used in experiments. These queries use all information in the original TIPSTER topics. In an on-line information retrieval environment, it is unrealistic to ask users to submit queries of that type. In this section, the experiments are designed to simulate an on-line retrieval environment to test how well an association thesaurus works. Here the description fields of the TIPSTER queries are taken as user query in an on-line environment. These user queries are used as natural language queries for the TIPFULL thesaurus, and the retrieved associated phrases are automatically added into queries. In the expanded queries, the duplicate phrases are added with weight 2.0 and the top ten nonduplicates with weight 1.0. Table 6 shows the experimental results, in which a 32.7% improvement is gained on the tip database.

	Precision (% change) – 50 queries						
Recall	INQ013	DUPONLY		NODUP		BOTH	
0	58.1	60.9	(+4.7)	68.3	(+17.5)	67.6	(+16.3)
10	30.9	39.1	(+26.5)	34.9	(+12.9)	41.4	(+34.1)
20	24.5	31.7	(+29.4)	27.9	(+13.8)	33.7	(+37.6)
30	20.3	26.1	(+28.8)	23.0	(+13.4)	27.4	(+35.1)
40	16.8	22.1	(+32.2)	19.4	(+15.6)	22.9	(+36.5)
50	13.5	18.4	(+36.4)	16.3	(+20.6)	19.5	(+44.4)
60	11.0	15.1	(+37.1)	13.3	(+21.3)	16.5	(+50.2)
70	9.0	11.9	(+32.7)	10.9	(+22.0)	13.3	(+48.0)
80	6.8	8.8	(+28.0)	8.6	(+26.0)	10.3	(+50.0)
90	4.4	5.4	(+23.1)	5.5	(+25.7)	6.6	(+49.2)
100	1.3	1.5	(+17.3)	1.3	(−0.1)	1.5	(+22.4)
average	17.9	21.9	(+22.6)	20.9	(+16.7)	23.7	(+32.7)

Table 6: Performance of description-only queries on tip

These improvements on TIPSTER are larger than those obtained with NPL. This suggests that bigger collections may produce better association thesauri.

7 Conclusions and Discussion

These experiments show that it is possible and feasible to construct useful collection-dependent association thesauri automatically without relevance judgments. It seems that the larger the collection, the better the association thesaurus performs. The approach to the construction is realistic and general. An association thesaurus can be converted to be suitable for the framework of other retrieval systems. This kind of thesaurus looks very promising for information retrieval.

The approach embodied in PhraseFinder is general, and can be extended to many retrieval problems. The association data generated by PhraseFinder is an information synthesis over a given collection. For example, by defining the phrase rule as a company or person name, PhraseFinder would generate association data for company or personal information systems. For a given query, INQUERY would output a list of company or person names associated with the query.

Many questions remain in using association thesauri to do query expansion. In our experiments, the same number of nonduplicate phrases are added into all queries. Obviously this is not necessarily optimal. How many phrases or terms should be added depends on the queries and the associated phrases. For some queries many phrases can be added, whereas other may degrade with only a few additional phrases. It is still not clear how to determine this number for a given query. In experiments on the NPL collection, a weighting scheme is used for query expansion. How weights should assigned to duplicate and nonduplicate phrases is experimentally determined. It is not clear that it is possible to compute these weights automatically. It is in question whether weighting values for one set of queries can be applied to another set of queries. It is still a puzzle why added phrases have to be down-weighted on small collections. Fortunately, query expansion methods using an association thesaurus are straightforward on a large collection like the TIPSTER, even though they may not be optimal.

Although data filtering techniques for association data not only reduce the amount of association data but also enhance the performance of thesauri, these techniques are still ad hoc and premature. It would be better to do data filtering using association strength based on association frequencies and other statistical values. A comprehensive data filtering technique can be expected to do better and to reduce association data further.

Acknowledgments

This research was supported by the NSF Center for Intelligent Information Retrieval at the University of Massachusetts. Discussions about the effect of different words on representing the content of texts with Bob Krovetz inspired the experiments on word-based association thesauri. Dan Nachbar, Steve Harding, and John Broglio supported the PhraseFinder development and experiments.

References

[1] James P. Callan, and W. Bruce Croft, *An Evaluation of Query Processing Strategies Using the TIPSTER Collection*, SIGIR'93 347-355, 1993.

[2] Kenneth Church, *A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text,* In Proc. of the 2nd Conf. on Applied Natural Language Processing, 136-143,. 1988.

[3] W.Bruce Croft, Howard R. Turtle, and David D. Lewis, *The Use of Phrases and Structured Queries in Information Retrieval*, SIGIR'91, 1991

[4] Carolyn J. Crouch, *An Approach to the Automatic Construction of Global Thesauri*, IP&M, Vol.26(5), 629-640, 1990

[5] Carolyn J. Crouch and Bokyung Yang, *Experiments in Automatic Statistical Thesaurus Construction*, SIGIR 92, 77-88, 1992

[6] David A. Evans, Kimberly Ginther-Webster, Mary Hart, Robert G. Lefferts, and Ira A. Monarch, *Automatic Indexing Using Selective NLP and First-Order Thesauri*, RIAO 91, 624-643, 1991

[7] Donna Harman, *Overview of the First Text REtrieval Conference*, SIGIR'93, 36-47, 1993

[8] R. Krovetz, *Viewing Morphology as an Inference Process*, SIGIR'93, 191-202, 1993

[9] Jack Minker, Gerald A. Wilson, and Barbara H. Zimmerman, *An Evaluation of Query Expansion by the Addition of Clustered Terms for a Document Retrieval System*, IP&M, Vol.8, 329-348, 1972

[10] C.J. Rijsbergen, D.J. Harper, and M.F. Porter, *The Selection of Good Search Terms*, IP&M, Vol.17, 77-91, 1981

[11] Gerda Ruge, *Experiments on Linguistically Based Term Associations*, RIAO 91, 528-545, 1991

[12] Porter M, *An Algorithm for Suffix Stripping*, Program, Vol.14(3), 130-137, 1980

[13] Yonggang Qiu, H.P. Frei, *Concept Based Query Expansion*, SIGIR'93, 160-169, 1993

[14] G. Salton, *Automatic Information Organization and Retrieval*, McGraw-Hill Book Company, 1968

[15] G. Salton, *Automatic Term Class Construction Using Relevance - A Summary of Word in Automatic Pseudoclassification*, IP&M, Vol.16 (1), 1-15, 1980

[16] G. Salton, C. Buckley, and C.T. Yu, *An Evaluation of Term Dependence Models in Information Retrieval*, LNCS 146, 151-173, 1983

[17] Alan F. Smeaton, *The Retrieval Effects of Query Expansion on a Feedback Document Retrieval System*, TR-2, Dept. of Computer Science, University College Dublin, 1982

[18] K. Spark Jones and R. M. Needham, *Automatic Term Classification and Retrieval*, IP&M, Vol.4, 91-100, 1968

[19] K. Spark Jones and D.M. Jackson, *The Use of Automatically-Obtained Keyword Classifications for Information Retrieval*, IP&M, Vol.5, 175-201, 1970

[20] K. Spark Jones, *Automatic Keyword Classification for Information Retrieval*, Butterworth, 1971.

[21] Howard R. Turtle, *Inference Networks for Document Retrieval*, Ph.D. Thesis, COINS Technical Report 90-92, University of Massachusetts at Amherst, 1990

[22] C. T. Yu, C. Buckley, K. Lam, and G. Salton, *A Generalized Term Dependence Model in Information Retrieval*, Information Technology: Research and Development, Vol.2, 129-154, 1983

[23] C.T. Yu, W. Meng, and S. Park, *A Framework for Effective Retrieval*, ACM Tran. on Database Systems, Vol.14,No.2, 147-167, 1989