PERGAMON

Information Processing and Management 36 (2000) 361–378

INFORMATION PROCESSING & MANAGEMENT

www.elsevier.com/locate/infoproman

Query expansion using heterogeneous thesauri

Rila Mandala*, Takenobu Tokunaga, Hozumi Tanaka

Department of Computer Science, Tokyo Institute of Technology, 2-12-1 Ookayama Meguro, Tokyo, 152-8552, Japan

Received 23 February 1999; accepted 27 October 1999

Abstract

This paper proposes a method to improve the performance of information retrieval systems by expanding queries using heterogeneous thesauri. The expansion terms are taken from a hand-crafted thesaurus, co-occurrence-based automatically constructed thesaurus, and predicate-argument-based automatically constructed thesaurus. To avoid the effects of wrong expansion terms, a weighting method is devised such that the weight of expansion terms depend not only on the weight of all terms in query, but also the weight of those terms in each thesaurus. Experiments show that using heterogeneous thesauri with an appropriate weighting method results in better retrieval performance than using only one type of thesaurus.

Keywords: Query expansion; Thesauri; Text retrieval

1. Introduction

One major problem in information retrieval is the difficulty in describing user information needs in terms of a query so that a system can accurately distinguish between relevant and irrelevant documents for the query (Salton & McGill, 1983). In fact, the user's original query statement will usually consist of just a few terms related to the subject of interest.

Query expansion is a technique utilized within information retrieval to remedy this problem (Ekmekcioglu, 1992; Fox, 1980). A query is expanded by adding other terms that are closely related to the original query terms. Expansion terms can be selected by referring to thesauri (Crouch, 1990; Crouch & Yang, 1992; Jing & Croft, 1994; Kristensen, 1993; Paice, 1991) or by

* Corresponding author. Tel.: +81-3-5734-2831; fax: +81-3-5734-2915.
E-mail address: rila@cs.titech.ac.jp (R. Mandala).

0306-4573/00/$ - see front matter
PII: S0306-4573(99)00068-0

consulting users through the relevance feedback technique (Salton & Buckley, 1990; Buckley & Salton, 1994). Past research has verified the effectiveness of relevance feedback, but it puts a burden on users to a certain extent. Furthermore, if a user is not familiar with the vocabulary of a document collection, it is difficult to obtain good expansion terms, unless the system can suggest terms to the user.

Currently, many information retrieval researchers use pseudo relevance feedback, which is relevance feedback without user intervention (Buckley & Salton, 1995; Xu and Croft, 1996). Firstly, several documents are retrieved as a result of the initial retrieval. Assuming that the top-*n* retrieved documents are relevant, the system uses the terms contained in those documents as expansion terms and retrieves again. This technique works well when the system has a fairly good performance and the top-*n* retrieved documents are really relevant. If this assumption does not hold, however, the performance of the system can degrade as the expansion terms are taken from irrelevant documents.

Query expansion using thesauri is still worth investigating. Briefly, there are two types of thesauri, that is, hand-crafted thesauri (Fox, 1980) and automatically constructed thesauri (Chen, Schatz, Yim & Fye, 1995; Crouch, 1990; Crouch & Yang, 1992). WordNet (Miller, Beckwidth, Fellbaum, Gross & Miller, 1990) is an example of a hand-crafted thesaurus which is available in machine readable form. Many researchers have tried to use relations defined in WordNet for query expansion. Unfortunately, the results have not been as good as expected. Voorhees (1994) used WordNet as a tool for query expansion. She conducted experiments using the TREC collection in which all terms in the queries were expanded using a combination of synonyms, hypernyms, and hyponyms. She set the weight of the words contained in the original query to 1, and used a combination of 0.1, 0.3, 0.5, 1, and 2 for the expansion terms. She then used the SMART information retrieval system (Salton, 1971) to retrieve the documents. Through her experiments, Voorhees succeeded in improving the system performance on only short queries with little or no significant improvement for long queries. She further tried to use WordNet as a tool for word sense disambiguation (Voorhees, 1993) and applied it to text retrieval, but the performance of retrieval was degraded.

Smeaton, Kelledy and O'Donnel (1996) tried to expand the queries of the TREC-4 collection with various strategies of weighting expansion terms, along with manual and automatic word sense disambiguation techniques. Unfortunately, all strategies degraded the retrieval performance.

Unlike hand-crafted thesauri, corpus-based thesauri are constructed automatically from corpora without human intervention. There are two different methods of extracting thesaural relationships from corpora, that is, co-occurrence statistics (Chen et al., 1995; Crouch, 1990; Qiu & Frei, 1993; Schutze & Pedersen, 1994) and grammatical relations (Grefenstette, 1992, 1994; Hindle, 1990; Jing & Croft, 1994; Ruge, 1992).

Qiu and Frei (1993) used an automatically constructed thesaurus and improved retrieval effectiveness by about 20% using small test collections. Schutze and Pedersen (1994, 1997) also built a co-occurrence-based thesaurus and applied it to two information retrieval applications. Using a scaled-down TREC collection, he slightly improved the retrieval performance.

Peat and Willett (1991) have provided theoretical evidence of the limitation of term co-occurrence data for query expansion in information retrieval. Consequently, some researchers have tried to build thesauri using more linguistically motivated methods. Ruge (1992) built a linguistically-based thesaurus, but she did not apply it to information retrieval. Grefenstette (1992) built a thesaurus using syntactic context and performed experiments using several small test collections. His method improved the performance for some small collections, but failed to improve the performance using other collections (Grefenstette, 1994). Jing and Croft (1994) also found an improvement through query expansion by using a grammatically-based automatically constructed thesaurus.

Despite the success of using corpus-based thesauri (Qiu & Frei, 1993; Jing & Croft, 1994; Schutze & Pedersen, 1997), the improvement is very limited when compared to the relevance-feedback method. This paper tries to analyze why query expansion using a thesaurus shows only limited performance, and based on this analysis we propose a new method to improve the performance of query expansion by combining heterogeneous thesauri for query expansion.

2. Analysis of thesaurus characteristics

As stated earlier, query expansion using WordNet has not always succeeded in improving the performance of information retrieval systems. One reason for this would be that WordNet is a general-purpose thesaurus. It has been compiled to be used in a wide variety of domains, and therefore lacks domain-specific thesaural relationships. We conducted query expansion experiments using WordNet and found limitations, which are summarized as follows:

- Two interrelated words may have different parts of speech. This is the case with "stochastic" (adjective) and "statistics" (noun). Since words in WordNet are grouped on the basis of part of speech, it is not possible to find a relationship between terms with different parts of speech.
- Most domain-specific relationships between two words are not found in WordNet as mentioned above.
- Some kinds of words, such as proper names, are not included in WordNet.

On the other hand, corpus-based thesauri are able to incorporate domain specific information, since they can be automatically constructed from a set of documents relating to a certain domain. Some researchers have succeeded in improving the performance of information retrieval systems with query expansion using corpus-based thesauri (Qiu & Frei, 1993; Jing & Croft, 1994; Schutze & Pedersen, 1997).

The construction of thesauri generally takes the following form (Charniak, 1993):

1. Extract word co-occurrences.
2. Define similarities (distances) between words on the basis of co-occurrences.
3. Cluster words on the basis of similarities.

Depending on the context of "co-occurrence", different types of thesauri can be constructed. In past attempts, document co-occurrence relations and predicate-argument relations have often been used as the co-occurrence context.

A drawback of using a document co-occurrence relation is that it is difficult to capture relations between certain words that share the same meaning. For example, "tumor" and "tumour" denote the same concept, but would never appear in the same document, at least not frequently enough so as to be recognized by a statistical method. This is because one tends not to use a variety of words denoting the same meaning in a single document. This kind of relationship can be found in a general purpose thesaurus such as WordNet, and can also be found in a thesaurus constructed using predicate-argument relations, since we can expect that "tumor" and "tumour" appear as an argument of the same predicate.

Since each type of thesaurus has different advantages and disadvantages, combining them provides a valuable tool for query expansion. In this paper, for the purpose of query expansion, we combine three types of thesauri: a hand-crafted general purpose thesaurus (WordNet), an automatically constructed thesaurus based on a document co-occurrence relation (co-occurrence-based thesaurus) and an automatically constructed thesaurus based on a predicate-argument relation (predicate-argument-based thesaurus).

3. Method

In this section, we first describe the construction method for each type of thesaurus utilized in this research, and then describe a term weighting method using similarity measures based on these thesauri.

3.1. WordNet

WordNet is a hand-crafted thesaurus developed by a Princeton University group led by George Miller (Miller et al., 1990). In WordNet, words are organized into taxonomies where each node is a set of synonyms (a "synset") representing a single sense. There are four different taxonomies based on different parts of speech and also there are many relationships defined among them. In our experiments we use only a noun taxonomy with hyponymy/hypernymy relations (or an *is–a* relation). Fig. 1 is a fragment of the WordNet taxonomy, where *is–a* relations are represented by the $\Rightarrow$ symbol.

The similarity between words a and b can be defined as the shortest path from each sense of a to each sense of b, as below (Resnik, 1995):

$$\mathrm{sim}_{ab} = \max\left[-\log\left(\frac{N_p}{2D}\right)\right]$$

```
tumor, tumour, neoplasm
       => growth
            => illness, malady, sickness
                 => ill health, unhealthiness, health problem
                      => physiological state
                           => condition, status
                                => state
```

Fig. 1. A fragment of WordNet taxonomy.

where N_p is the number of nodes in path p from a to b and D is the maximum depth of the taxonomy. We normalize the similarity value so that its value ranges from 0 to 1.

3.2. Co-occurrence based thesaurus

This method is based on the assumption that a pair of words that occur frequently together in the same document are related to the same subject. Therefore, word co-occurrence information can be used to identify semantic relationships between words. Suppose two words a and b occur in df_a and df_b documents, respectively, and co-occur in df_c documents, then the similarity between a and b can be calculated using a similarity coefficient such as the Dice coefficient (Peat & Willett, 1991)

$$\frac{2 \times df_c}{df_a + df_b}.$$

3.3. Predicate-argument-based thesaurus

In this method term, relations are gathered on the basis of linguistic relations and not document co-occurrence statistics (Hindle, 1990). Words appearing in a similar grammatical context are assumed to be similar, and, therefore, classified into the same class. As grammatical relations, we use subject–verb, verb–object and adjective–noun relations in order to build word classes.

First, all the documents are parsed using the Apple Pie Parser, which is a probabilistic chart parser developed by Sekine (Sekine & Grishman, 1995). Then the following syntactic structures are extracted.

- Subject–verb.
- Verb–object.
- Adjective–noun.

Each noun has a set of co-occurring verbs and adjectives, and for each such relationship, a Dice coefficient value is calculated.

- $C_{\mathrm{sbj}}(v_i, n_j) = [2 \times f_{\mathrm{sbj}}(v_i, n_j)/f(v_i) + f_{\mathrm{sbj}}(n_j)]$, where $f_{\mathrm{sbj}}(v_i, n_j)$ is the frequency of noun n_j occurring as the subject of verb v_i, $f_{\mathrm{sbj}}(n_j)$ is the frequency of noun n_j occurring as the subject of any verb, and $f(v_i)$ is the frequency of verb v_i.
- $C_{\mathrm{obj}}(v_i, n_j) = [2 \times f_{\mathrm{obj}}(v_i, n_j)/f(v_i) + f_{\mathrm{obj}}(n_j)]$, where $f_{\mathrm{obj}}(v_i, n_j)$ is the frequency of noun n_j occurring as the object of verb v_i, $f_{\mathrm{obj}}(n_j)$ is the frequency of noun n_j occurring as the object of any verb, and $f(v_i)$ is the frequency of verb v_i.
- $C_{\mathrm{adj}}(a_i, n_j) = [2 \times f_{\mathrm{adj}}(a_i, n_j)/f(a_i) + f_{\mathrm{adj}}(n_j)]$, where $f_{\mathrm{adj}}(a_i, n_j)$ is the frequency of noun n_j being modified by adjective a_i, $f_{\mathrm{adj}}(n_j)$ is the frequency of noun n_j being modified by any adjective, and $f(a_i)$ is the frequency of adjective a_i.

We define the similarity of two nouns with respect to a predicate, as the minimum of the following Dice coefficients:

$$\mathrm{sim}_{\mathrm{sbj}}(v_i, n_j, n_k) = \min\{C_{\mathrm{sbj}}(v_i, n_j), C_{\mathrm{sbj}}(v_i, n_k)\}$$

$$\mathrm{sim}_{\mathrm{obj}}(v_i, n_j, n_k) = \min\{C_{\mathrm{obj}}(v_i, n_j), C_{\mathrm{obj}}(v_i, n_k)\}$$

$$\mathrm{sim}_{\mathrm{adj}}(a_i, n_j, n_k) = \min\{C_{\mathrm{adj}}(a_i, n_j), C_{\mathrm{adj}}(a_i, n_k)\}.$$

Finally, overall similarity between two nouns is defined as the average of all the similarities between those two nouns for all predicate-argument structures.

3.4. Expansion term weighting method

A query q is represented in terms of a vector $\mathbf{q} = (w_1, w_2, \ldots, w_n)$, where n is the total number of terms in the collection, w_i is the weight of term t_i when t_i is contained in query q. If t_i does not appear in the query, w_i becomes 0.

The similarity between a query q and term t_j can be defined as follows (Qiu & Frei, 1993):

$$\mathrm{sim}_{qt}(q, t_j) = \sum_{t_i \in q} w_i \cdot \mathrm{sim}(t_i, t_j)$$

where $\mathrm{sim}(t_i, t_j)$ is defined as the average of the similarities of the three types of thesauri mentioned in the previous section.

With respect to a query q, all the terms in a collection can be ranked according to their sim_{qt}. Terms t_j with high sim_{qt} in rank are used as expansion terms.

The weight $w_{\mathrm{ex}}(q, t_j)$ of an expansion term t_j with respect to a query q is defined as the following function of $\mathrm{sim}_{qt}(q, t_j)$:

$$w_{\mathrm{ex}}(q, t_j) = \frac{\mathrm{sim}_{qt}(q, t_j)}{\sum_{t_i \in q} w_i}.$$

The weight of an expansion term depends both on all terms appearing in a query and on the similarity between the terms, and ranges from 0 to 1. This weight can be interpreted mathematically as the weighted mean of similarities between term t_j and all terms in the query. The weight of the original query terms are weighting factors of those similarities.

A query q is expanded by adding the vector

$$\mathbf{q}_e = (w'_1, w'_2, \ldots, w'_n)$$

to the original query vector, where w'_j is equal to $w_{\mathrm{ex}}(q, t_j)$ if t_j is in the top r ranked terms, otherwise w'_j becomes 0.

4. Experiments

In order to evaluate the effectiveness of the proposed method, we conducted experiments using the TREC-7 information retrieval test collection (Voorhees & Harman, 1999). The

TREC-7 test collection consists of 50 topics (queries) and 528,155 documents from several sources: the Financial Times (FT), Federal Register (FR94), Foreign Broadcast Information Service (FBIS) and the LA Times. Each topic consists of three sections, the "Title", "Description" and "Narrative". Table 1 shows statistics of the document collection, Table 2 shows statistics of the topics, and Fig. 2 shows an example of the topic.

For the query, we use, respectively, the title only, the description only, and all sections of the topic. Note that in the TREC-7 collection, the description section contains all the terms in the title section.

As a baseline we used the SMART version 11.0 (Salton, 1971) system without query expansion. SMART is an information retrieval engine based on the vector space model in which term weights are calculated based on term frequency, inverse document frequency, and document length normalization. The weighting method for document collection is as follows:

$$\frac{(\log(tf_{ik}) + 1.0)}{\sqrt{\sum_{j=1}^{n}[\log(tf_{ij} + 1.0)]^2}}$$

and the weighting method for the initial query is as follows:

$$\frac{(\log(tf_{ik}) + 1.0) * \log(N/n_k)}{\sqrt{\sum_{j=1}^{n}[\log(tf_{ij} + 1.0) * \log(N/n_j)]^2}}$$

where tf_{ik} is the occurrence frequency of term t_k in query q_i (for query terms weighting) or in document d_i (for document terms weighting), N is the total number of documents in the collection, and n_k is the number of documents to which term t_k is assigned. This term weighting method is called the *lnc.ltc* weighting method.

The results are shown in Table 3. This table shows the average of non-interpolated precision for each case, expansion using only WordNet, expansion using only the predicate-argument based thesaurus, expansion using only the co-occurrence based thesaurus, and expansion using the combination of all thesauri types. For each method we give the percentage of improvement over the baseline method in parentheses. We demonstrate that the performance using the

Table 1
TREC-7 document statistics

Source	Size (MB)	No. of docs	words/doc. (median)	words/doc. (mean)
Disk 4				
FT	564	210,158	316	412.7
FR94	395	55,630	588	644.7
Disk 5				
FBIS	470	130,471	322	543.6
LA Times	475	131,896	351	526.5

Table 2
TREC-7 topic length statistics (words)

Topic section	Min.	Max.	Mean
Title	1	3	2.5
Description	5	34	14.3
Narrative	14	92	40.8
All	31	114	57.6

combined thesauri for query expansion is better than SMART version 11.0 using the *lnc.ltc* term weighting method without expansion and than expansion using just one type of thesaurus.

5. Analysis of results

5.1. Contribution of each thesaurus

We investigated to what extent each thesaurus contributes in providing expansion terms. Table 4 summarizes the percentage of expansion terms in expanded queries that are suggested by one thesaurus only, two of the thesauri, and all thesauri. We can see that each thesaurus contributes almost the same in providing expansion terms. Table 5 shows example expansion terms and their source thesauri, for the query: *how often were the peace talks in Ireland delayed or disrupted as a result of acts of violence?*

5.2. Different coefficient measures

We also investigated the effect of different coefficient measures for constructing thesauri.

Title
commercial cyanide uses

Description
What are the industrial or commercial uses of cyanide or its derivatives?

Narrative
A document is relevant if it names or describes a process that uses cyanide commercially or mentions that cyanide-rich waste comes from a particular industry.

Fig. 2. Topic example.

Table 3
A comparison of the average non-interpolated precision for baseline, single, pairwise, and combined thesauri

Topic type	Base	Expanded with						
		WordNet only	Pred-arg only	Co-occur only	WordNet + pred-arg	WordNet + co-occur	Pred-arg + co-occur	Combined method
Title	0.1175	0.1276 (+8.6%)	0.1473 (+25.4%)	0.1603 (+36.4%)	0.1611 (+37.1%)	0.1698 (+44.5%)	0.1859 (+58.2%)	0.2213 (+88.3%)
Description	0.1428	0.1509 (+5.7%)	0.1626 (+17.1%)	0.1846 (+29.3%)	0.1832 (+28.3%)	0.1973 (+38.2%)	0.2315 (+62.1%)	0.2590 (+81.4%)
All	0.1976	0.2010 (+1.7%)	0.2203 (+11.5%)	0.2333 (+18.1%)	0.2276 (+15.2%)	0.2423 (+22.6%)	0.2565 (+29.8%)	0.2573 (+30.2%)

- Mutual information measure (MI)

  The mutual information measure compares the probability of the co-occurrence of words a and b with the probabilities of the occurrence of a and b independently (Hindle, 1990). The definition of the mutual information measures is as follows:

$$\mathrm{MI}(a, b) = \log \frac{P(a, b)}{P(a) \cdot P(b)}$$

  The probabilities of $P(a)$ and $P(b)$ are estimated by counting the number of occurrences of a and b in document collection and normalizing by the total number of words in the document collection. The value of $P(a, b)$ is estimated by counting the number of times that word a co-occurs with b and is also normalized.
- Tanimoto coefficient

  Suppose two words a and b occur in df_a and df_b documents, respectively, and co-occur in df_c documents, then the similarity of a and b can be calculated using the Tanimoto coefficient (Peat & Willett, 1991) as follows:

$$\frac{df_c}{df_a + df_b - df_c}.$$

Fig. 3 shows a recall-precision graph using the description section of the topics, and Table 6 shows the average of non-interpolated precision for different topic sections using different similarity measures. Although the mutual information measure gives better performance than the other coefficient measure methods, the difference is not significant.

5.3. *Number of expansion terms*

Fig. 4 shows the effect of the number of expansion terms added to the original query, on retrieval performance. The retrieval performance is measured by the average non-interpolated precision over all relevant documents. The improvement in expansion increases when adding up to 20 terms, and reaches a plateau, then the improvement begins to decrease when more than 50 terms are added. The expanded queries, however, still perform better than the original queries. The results shown in Fig. 4 also indicate that expanding a query with 30 to 40 top ranked terms seems to be the safest method with respect to the collection targeted in this evaluation.

Table 4
Percentage of expansion terms in the expanded query suggested by given source thesauri

Source thesauri	Title (%)	Description (%)	All (%)
WordNet	6	7	5
Co-occurrence	12	9	20
Pred-arg	10	8	14
WordNet and co-occurrence	14	13	9
WordNet and pred-arg	13	13	7
Co-occurrence and pred-arg	17	28	32
All	28	22	13

5.4. *Size of corpora for thesaurus construction*

We now describe an experiment to see the effect on retrieval performance of the size of corpora which thesauri are constructed from. In other words, we would like to see whether the combined thesaurus needs to be updated when new documents are added to the collection. We first built thesauri using only TREC-7 disk 4, and performed retrieval experiments using disks 4 and 5. Figs. 5–7 show the improvement in retrieval performance with expanded queries using these thesauri over the original queries using the title only, the description only, and all parts of the topic, respectively. For comparison, we also show the improvement using the combination of thesauri from the entire collection. The results indicate that despite only half the collection being used to build thesauri, the performance is not too degraded. In fact, we still get a significant improvement over the original queries. This means that the combined set of thesauri constructed from the sample collection is still useful for the entire collection.

5.5. *Expansion using only thesauri intersection*

We compare the retrieval results of using our method described above with the results of

Table 5
Example expansion term and their source thesauri

Source thesauri	Expansion terms
WordNet	Force
Co-occurrence	Uup, RUC, Downing
Pred-arg	Anglo-Irish
WordNet and co-occurrence	Treaty, paramilitary
WordNet and pred-arg	Irish, British, declaration
Co-occurrence and pred-arg	Sinn, Fein, Gerry, Adams, Heathrow, John, Major, Albert, Reynolds, Crossmaglen
All	IRA, protestant, catholic, politics, meeting, bomb, terrorism, Dublin, street, ceasefire, firebomb, gun, incident, unionist, constabulary, loyalist, crime, campaign, killing, arm, fire, explosion

expansion using only words that appear in the intersection of the suggestions made by all three thesauri. Table 7 shows the non-interpolated average precision of SMART with the *lnc.ltc* weighting method without expansion, our method, and expansion using only thesauri intersection. As we can see in Table 7, our method gives significantly better results than using only the intersection of the thesauri.

5.6. Comparison with relevance-feedback

We compare our method with query expansion using the relevance feedback technique, in which 30 documents among the documents retrieved in the initial retrieval are used for feedback. We use the Rocchio formula for term reweighting as follows:

$$Q_{\text{new}} = \alpha \cdot Q_{\text{old}} + \beta \sum_{r=1}^{n_{\text{rel}}} \frac{D_r}{n_{\text{rel}}} - \gamma \sum_{n=1}^{n_{\text{nonrel}}} \frac{D_n}{n_{\text{nonrel}}}$$

where α, β and γ are constants, D_r is the vector of a relevant document d_r, D_n is the vector of an irrelevant document d_n, n_{rel} is the number of relevant documents retrieved, and n_{nonrel} is the number of irrelevant documents. We set $\alpha = 8$, $\beta = 16$, and $\gamma = 4$ for this experiment (Buckley & Salton, 1994, 1995).

For ideal relevance feedback, 30 relevant documents are determined by referring to the relevance judgement on the test collection, and for the pseudo relevance feedback, the 30 top-ranked documents in the initial retrieval are assumed relevant.

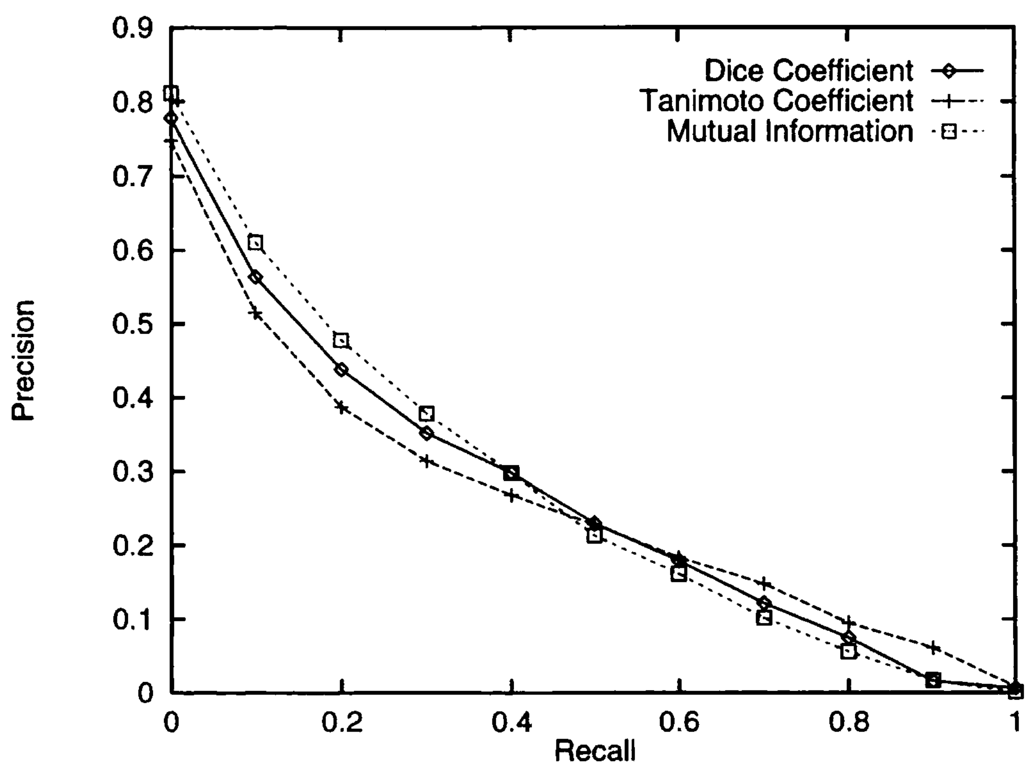

Fig. 3. Eleven-point precision for different similarity measures.

Table 6
Comparison of different similarity measures for different topic sections

Topic section	Base	Dice coefficient	MI	Tanimoto coefficient
Title	0.1175	0.2213 (+88.3%)	0.2314 (+96.9%)	0.2011 (+71.7%)
Description	0.1428	0.2590 (+81.4%)	0.2645 (+85.2%)	0.2492 (+75.3%)
All	0.1976	0.2573 (+30.2%)	0.2724 (+37.8%)	0.2654 (+34.5%)

Figs. 8–10 show 11-point recall-precision curves comparing our method, ideal relevance feedback, and pseudo relevance feedback using the title only, the description only, and all parts of the topic, respectively. It shows that our method significantly outperform pseudo relevance feedback and is slightly worse than ideal relevance feedback.

6. Discussion

The key techniques used in our method can be summarized as follows:

- Broadening thesaurus coverage by combining several thesauri.
- Weighting expansion terms to avoid wrong expansion.

The three types of thesauri used have different characteristics. Automatically constructed

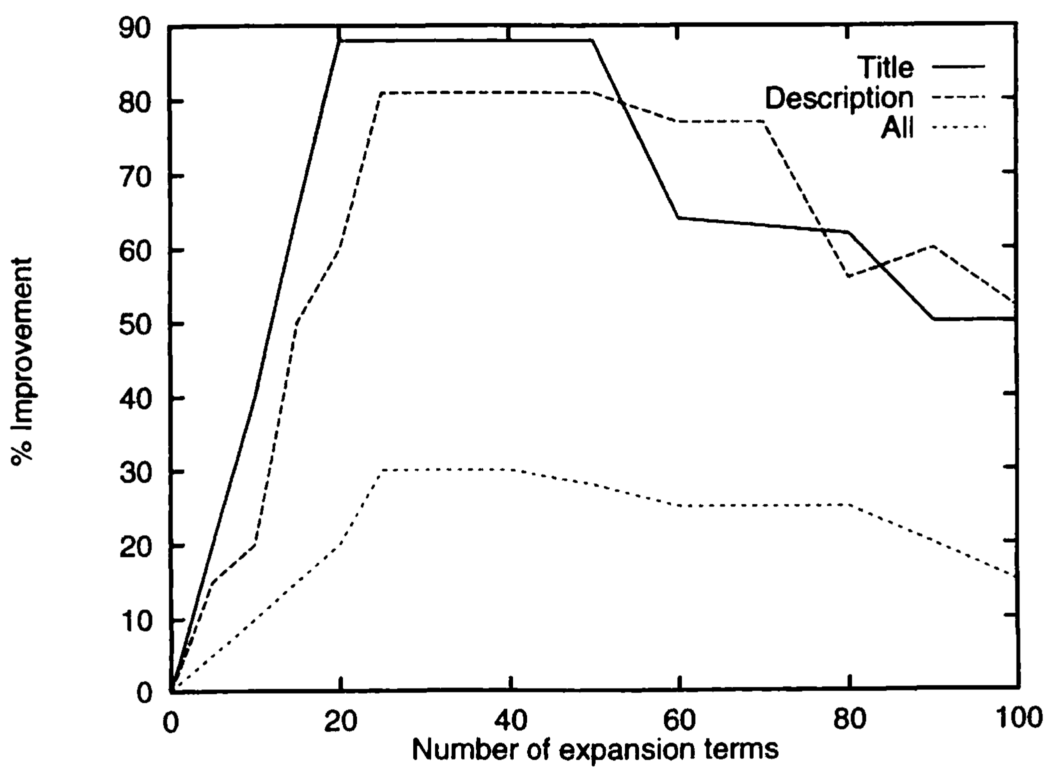

Fig. 4. Relation between the number of expansion terms and improvement.

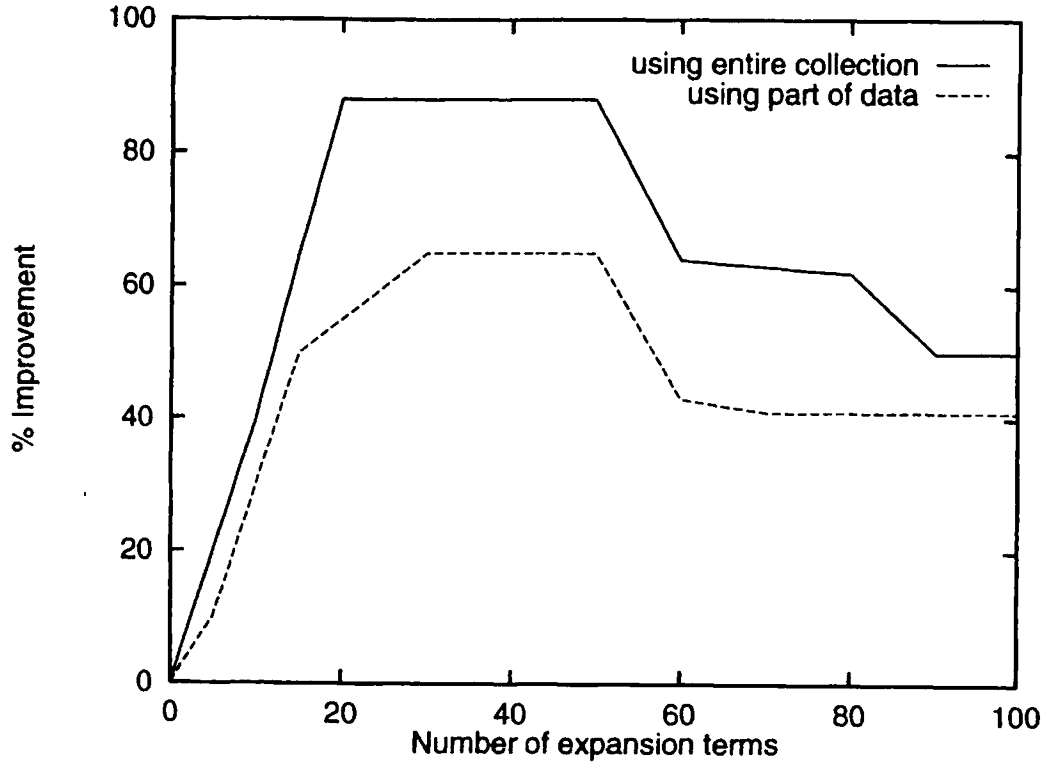

Fig. 5. Relation between the proportion of the corpus used and relative system improvement (title).

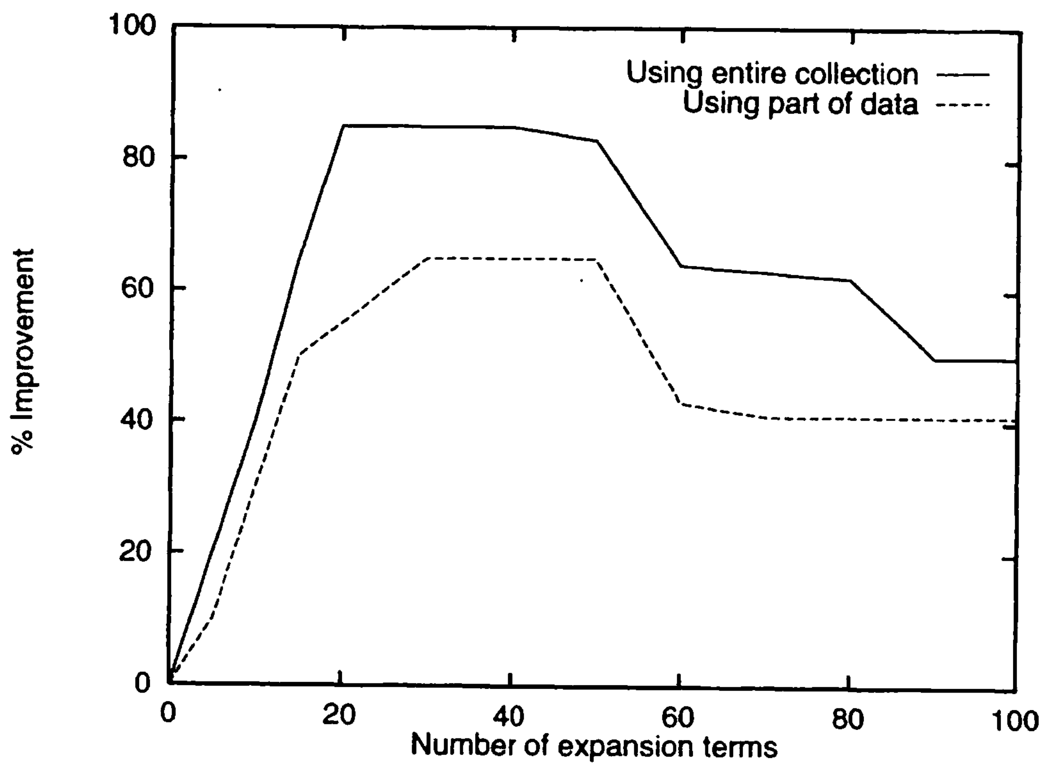

Fig. 6. Relation between the proportion of the corpus used and relative system improvement (description).

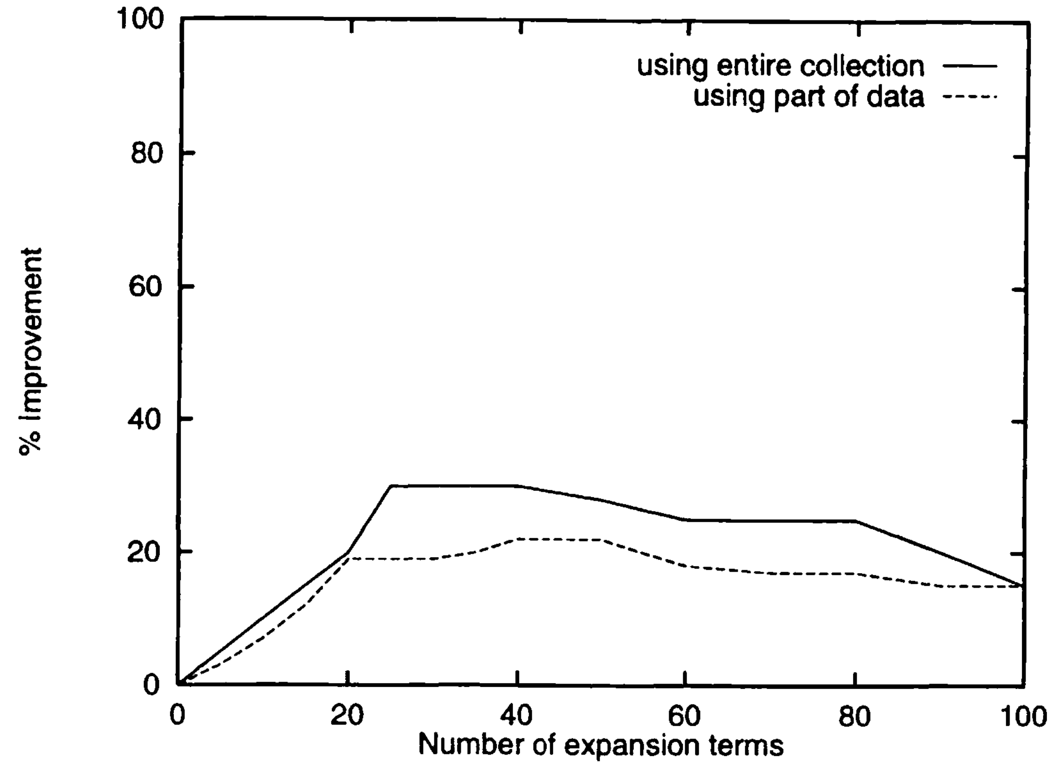

Fig. 7. Relation between the proportion of the corpus used and relative system improvement (all).

thesauri add not only new terms but also new relationships not found in WordNet. If two terms often co-occur in a document then those two terms are likely to bear some relationship. Why not only use the automatically constructed thesauri? The answer is that some general knowledge may be missing in automatically constructed thesauri.

The advantages of our weighting method can be summarized as follows:

- The weight of each expansion term considers the similarity of that term to all terms in the original query, rather than to just one query term.
- The weight of an expansion term also depends on its similarity in all types of thesauri.

This method can accommodate the polysemous word problem, because an expansion term taken from a different sense to the original query term sense is given very low weight. The reason for this is that the weighting method depends on all query terms and all of the thesauri. For example, the word "bank" has many senses in WordNet. Two such senses are a financial

Table 7
A comparison of the results of our method and using the intersection of the thesauri

Topic type	Without expansion	Our method	Intersection
Title	0.1175	0.2213 (+88.3%)	0.1553 (+32.1%)
Description	0.1428	0.2590 (+81.4%)	0.1789 (+25.3%)
All	0.1976	0.2573 (+30.2%)	0.2274 (+15.1%)

institution and river edge. In a document collection relating to financial banks, the river sense of "bank" will generally not be found in the co-occurrence based thesaurus because of a lack of documents talking about rivers. Even though (with small possibility) there may be some documents in the collection talking about rivers, if the query contained the finance sense of "bank" then the other terms in the query would also be concerned with finance and not rivers. Thus, rivers would only relate to the term "bank" and there would be no relationship with other terms in the original query, resulting in a low weight.

Our method only retrieves documents once and only needs extra time for computing the multiplication of original query terms weight and the similarity value between terms in the combined thesauri which are pre-computed, so that the time required by our method is proportional to the number of query terms. In contrast, relevance feedback needs to compute the similarity value between query terms and all documents in the second retrieval process, so that the time required by this method is proportional to the number of terms in all documents. Hence, our method requires less retrieval time than relevance feedback.

7. Conclusions and future works

We have proposed the use of heterogeneous thesauri for query expansion. The underlying idea is that each type of thesaurus has different characteristics and, therefore, their combination can provide a valuable resource for query expansion. Wrong expansion terms are

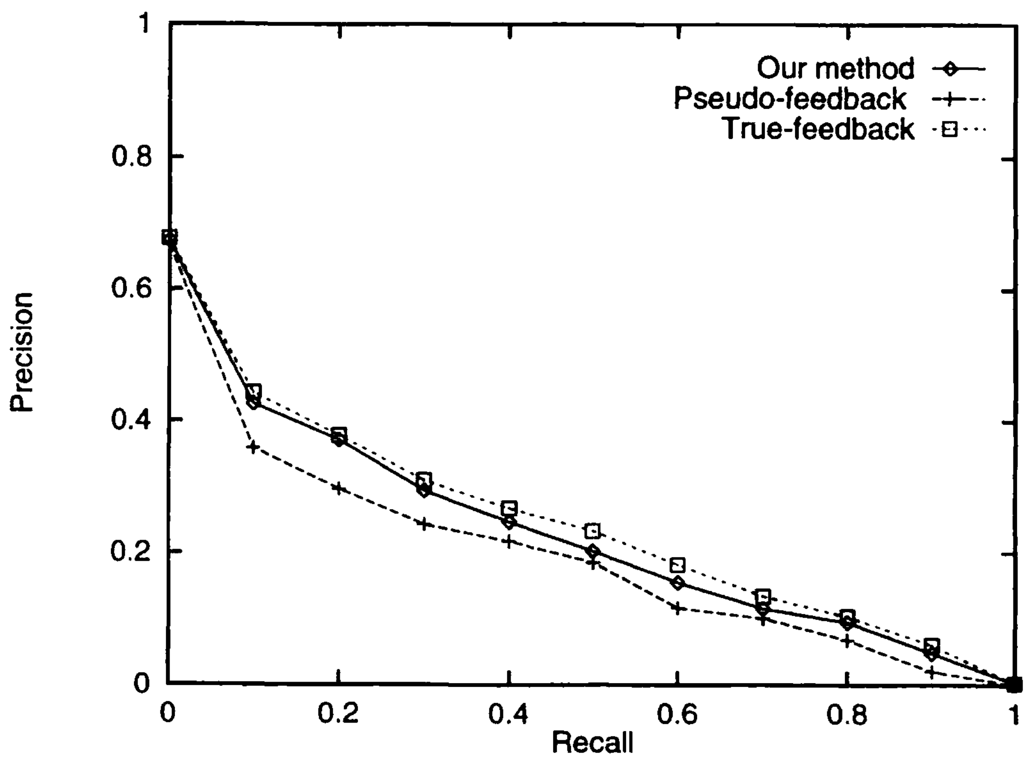

Fig. 8. Comparison of performance between our method and relevance feedback (title).

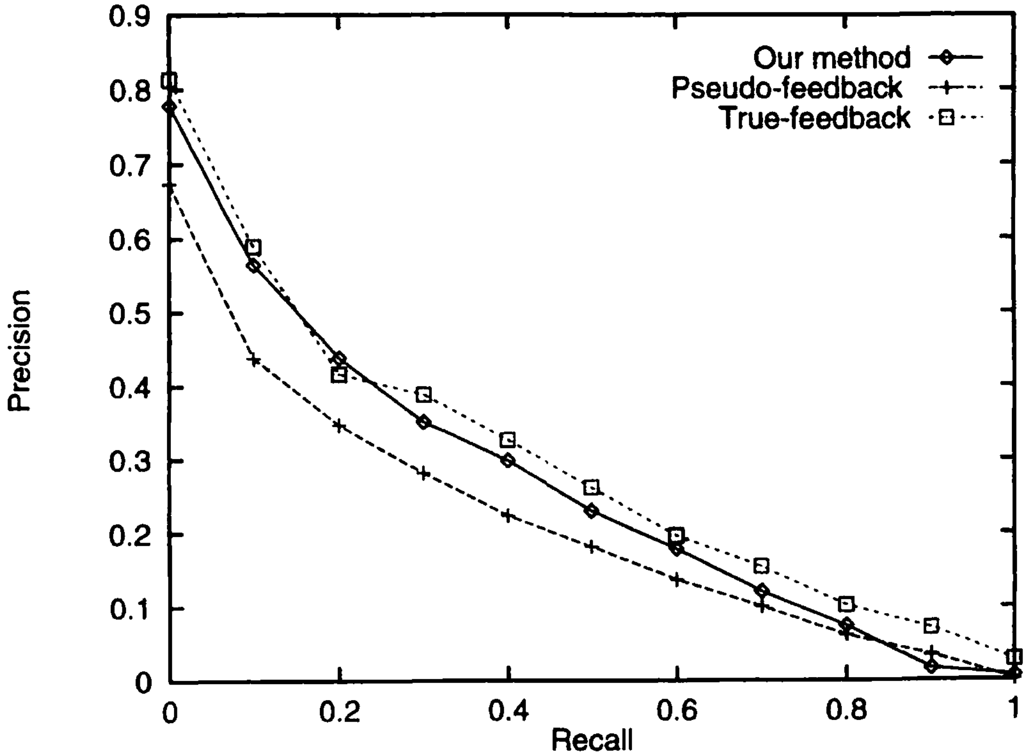

Fig. 9. Comparison of performance between our method and relevance feedback (description).

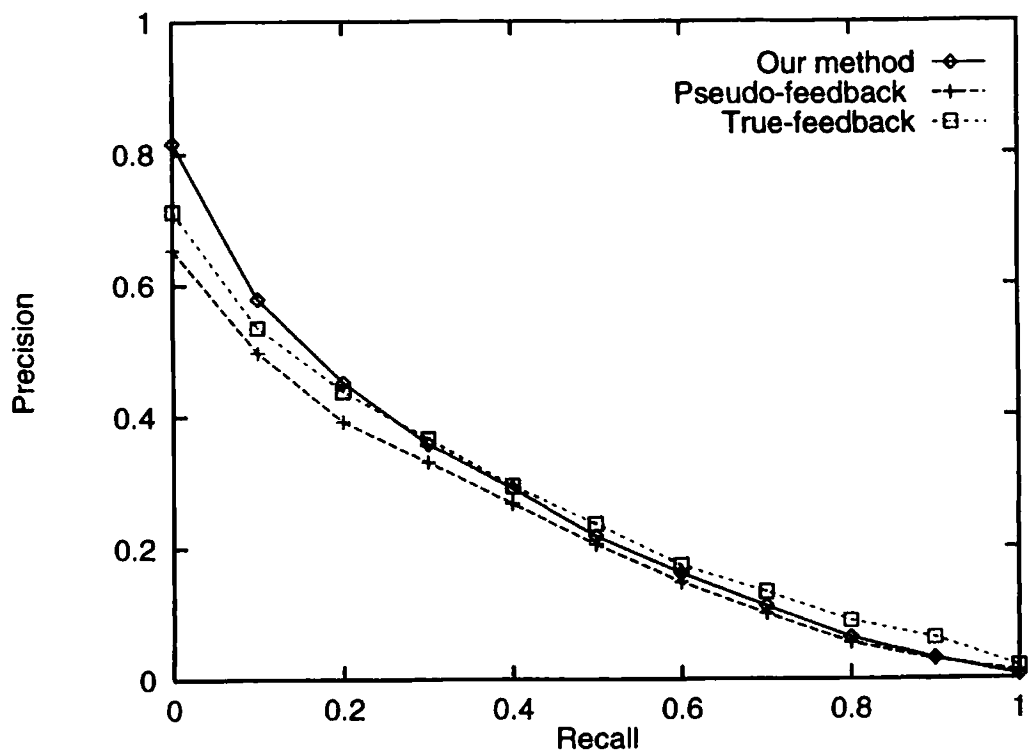

Fig. 10. Comparison of performance between our method and relevance feedback (all).

avoided by designing a weighting term method in which the weight of expansion terms not only depends on all query terms, but also on similarity measures in all types of thesauri. This is actually a kind of word sense disambiguation.

Experiments have shown that use of the combined set of thesauri gives better retrieval results than just using one type of thesaurus. Our method also out-performs pseudo relevance feedback, although more experiments are needed for further investigation.

Future research will include the use of parser with better performance and the use of anaphora resolution to accurately determine the nature of relationships involving proper names.

Acknowledgements

The authors would like to thank the anonymous referees and Timothy Baldwin for useful comments on the earlier version of this paper. We also thank Chris Buckley (SabIR Research) for support with SMART and Satoshi Sekine (New York University) for the Apple Pie Parser. This research is partially supported by JSPS project number JSPS-RFTF96P00502.

References

Buckley, C., & Salton, G. (1994). The effect of adding relevance information in a relevance feedback environment. In B. Croft, & C. van Rijsbergen, *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, Dublin, Ireland* (pp. 292–300). London. UK: Springer-Verlag 3–6 July.

Buckley C., & Salton G. (1995). Automatic query expansion using SMART: TREC-3. In D. Harman (Ed.). *Overview of the Third Text Retrieval Conference (TREC-3)* (pp. 69–80). Gaithersburg, MD: NIST Special Publication 500-225.

Charniak, E. (1993). *Statistical Language Learning*. Cambridge, MA: MIT Press.

Chen, H., Schatz, B., Yim, T., & Fye, D. (1995). Automatic thesaurus generation for an electronic community system. *Journal of the American Society for Information Science*, *46*(3), 175–193.

Crouch, C. J. (1990). An approach to the automatic construction of global thesauri. *Information Processing and Management*, *26*(5), 629–640.

Crouch, C., & Yang, B. (1992). Experiments in automatic statistical thesaurus construction. In N. Belkin, P. Ingwersen, & A. M. Pesjtersen, *Proceedings of the 15th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, Copenhagen, Denmark* (pp. 77–82). New York, NY: ACM Press 21–24 June.

Ekmekcioglu, F. (1992). Effectiveness of query expansion in ranked-output document retrieval systems. *Journal of Information Service*, *18*(2), 139–147.

Fox, E. A. (1980). Lexical relations enhancing effectiveness of information retrieval systems. *SIGIR Forum*, *15*(3), 6–36.

Grefenstette, G. (1992). Use of syntactic context to produce term association lists for text retrieval. In N. Belkin, P. Ingwersen, & A. M. Pesjtersen, *Proceedings of the 15th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, Copenhagen, Denmark* (pp. 89–97). New York, NY: ACM Press 21–24 June.

Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery*. Dordrecht, Netherlands: Kluwer.

Hindle, D. (1990). Noun classification from predicate-argument structures. In R. C. Berwick, *Proceedings of the 28th*

Annual Meeting of the Association for Computational Linguistics, Pittsburgh, USA (pp. 268–275). Morristown, NJ: Association for Computational Linguistics 6–9 June.

Jing, Y., & Croft, B. (1994). An association thesaurus for information retrieval. In F. Brentano, & F. Seitz, *Proceedings of the RIAO'94 Conference, New York, USA* (pp. 146–160). Paris, France: C.I.D.-C.A.S.I.S 11–13 October.

Kristensen, J. (1993). Expanding end-users query statements for free text searching with a search-aid thesaurus. *Information Processing and Management, 29*(6), 733–744.

Miller, G., Beckwidth, R., Fellbaum, C., Gross, D., & Miller, K. (1990). Wordnet: an on-line lexical database. *International Journal of Lexicography: Special Issue, 3*(4), 235–244.

Paice, C. D. (1991). A thesaural model for information retrieval. *Information Processing and Management, 27*(5), 433–447.

Peat, J., & Willett, P. (1991). The limitations of term co-occurrence data for query expansion in document retrieval systems. *Journal of the American Society for Information Science, 42*(5), 378–383.

Qiu, Y., & Frei, H. (1993). Concept based query expansion. In R. Korfhage, E. Rasmussen, & P. Willett, *Proceedings of the 16th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, Pittsburgh, USA* (pp. 160–169). New York, NY: ACM Press 27 June–1 July.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In C. Mellish, *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI-95), Montreal, Canada* (pp. 448–453). San Francisco, CA: Morgan Kaufmann 20–25 August.

Ruge, G. (1992). Experiments on linguistically-based term associations. *Information Processing and Management, 28*(3), 317–332.

Salton, G. (1971). *The SMART Retrieval System Experiments in Automatic Document Processing*. Englewood Cliffs, NJ: Prentice-Hall.

Salton, G., & Buckley, C. (1990). Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science, 41*(4), 288–297.

Salton, G., & McGill, M. (1983). *An Introduction to Modern Information Retrieval*. New York, NY: McGraw-Hill.

Schutze, H., & Pedersen, J. (1994). A cooccurrence-based thesaurus and two applications to information retrieval. In F. Brentano, & F. Seitz, *Proceedings of the RIAO'94 Conference, New York, USA* (pp. 266–274). Paris, France: C.I.D.-C.A.S.I.S 11–13 October.

Schutze, H., & Pederson, J. (1997). A cooccurrence-based thesaurus and two applications to information retrieval. *Information Processing and Management, 33*(3), 307–318.

Sekine, S., & Grishman, R. (1995). A corpus-based probabilistic grammar with only two non-terminals. In *Proceedings of the International Workshop on Parsing Technologies, Prague, Czech Republic*.

Smeaton, A., Kelledy, F., & O'Donnel, R. (1996). TREC-4 Experiments at Dublin City University: thresholding postings lists, query expansion by word–word distances and POS tagging of Spanish text. In D. Harman, *Proceedings of the Fourth Text Retrieval Conference (TREC-4)* (pp. 373–390) NIST Special Publication 500-236, Gaithersburg, MD.

Voorhees, E. (1993). Using WordNet to disambiguate word senses for text retrieval. In R. Korfhage, E. Rasmussen, & P. Willett, *Proceedings of the 16th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, Pittsburgh, USA* (pp. 171–180). New York, NY: ACM Press 27 June–1 July.

Voorhees, E. (1994). Query expansion using lexical-semantic relations. In B. Croft, & C. van Rijsbergen, *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, Dublin, Ireland* (pp. 61–69). London, UK: Springer-Verlag 3–6 July.

Voorhees, E., & Harman, D. (1999). Overview of the seventh text conference (TREC-7). In *Proceedings of the Seventh Text Retrieval Conference* NIST Special Publication 500-242.

Xu, J., & Croft, B. (1996). Query expansion using local and global document analysis. In H.-P. Frei, & P. Schauble, *Proceedings of the 19th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, Zurich, Switzerland* (pp. 4–11). New York, NY: ACM Press 18–22 August.