Combining Multiple Evidence from Different Types of Thesaurus for Query Expansion

Rila Mandala, Takenobu Tokunaga, and Hozumi Tanaka
Department of Computer Science
Tokyo Institute of Technology
2-12-1 Oookayama Meguro-Ku
Tokyo 152-8522 Japan
{rila,take,tanaka}@cs.titech.ac.jp

Abstract

Automatic query expansion has been known to be the most important method in overcoming the word mismatch problem in information retrieval. Thesauri have long been used by many researchers as a tool for query expansion. However only one type of thesaurus has generally been used. In this paper we analyze the characteristics of different thesaurus types and propose a method to combine them for query expansion. Experiments using the TREC collection proved the effectiveness of our method over those using one type of thesaurus.

1 Introduction

Document authors and users use a great variety of words to refer to the same concepts [21]. As such, information retrieval systems must bridge the semantic gap which exists between the vocabulary of authors and that of users.

Query expansion is one method to solve the above problem [4, 5]. Query expansion can be performed either manually or automatically. In this paper, we are concerned with automatic query expansion. Query expansion can take place prior to either the initial search or the relevance feedback search [20, 1]. Automatic relevance feedback expansion is generally implemented by adding words that occur in top-n ranked documents, but were not included in the original query. Expansion may involve all of the terms in relevant documents, or some subset of them.

In this paper we focus on expansion prior to the initial search, which can be achieved using a thesaurus. Thesauri have frequently been incorporated in information retrieval systems as a device for the recognition of synonymous expressions and linguistic entities that are semantically similar but superficially distinct.

Automatic query expansion using thesauri has been the target of research for nearly four decades, and a lot of methods have been proposed. The various methods can be classified into the following 3 basic groups :

1. Hand-crafted thesaurus based [27, 25].
2. Co-occurrence-based automatically constructed thesaurus based [2, 3, 16, 22].
3. Head-modifier-based automatically constructed thesaurus based [7, 6, 11, 12, 18].

Query expansion based on hand-crafted thesauri only succeeds if a domain-specific thesauri is used which corresponds closely to domain-specific document collection [5]. The use of general-purpose hand-crafted thesaurus for automatic query expansion has not been very successfull [27, 25]. Voorhees [27] 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 [19] 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 [26] and applied it to text retrieval, but the performance of retrieval was degraded.

Smeaton [25] 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. Qiu [16] used an automatically constructed thesaurus to improve retrieval effectiveness by about 20% using small test collections. Schutze [22, 23] 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 Willet [15] have provided theoretical evidence of the limitations 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 [18] built a linguistically-based thesaurus, but she did not apply it to information retrieval. Grefenstette [7] 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 [6]. Jing [12] also found an improvement

SIGIR '99 8/99 Berkley, CA USA

through query expansion by using a grammatically-based automatically constructed thesaurus.

After we analyze the characteristics of the three types of thesaurus given above, we propose a method to utilize the different types of thesaurus together, as evidence for query expansion.

2 Thesaurus Characteristics

General-purpose thesauri are not specific enough to offer synonyms for words as used in the corresponding document collection. For example, in a document related to computers, the word *bug* will have meanings quite differrent from everyday language. General-purpose thesauri also do not cover many words found in queries such as proper nouns, yet proper nouns are often a good retrieval indicator.

Co-occurrence-based thesauri can capture the domain-specific meanings of words, because they are constructed from the document collection. However it is difficult to determine the appropriate word window size within which to consier co-occurrence. For example, if the window size is one document then every word in the document is considered potentially related to every other word, no matter what the distance between them. When smaller windows are used, similar effects are still seen.

The second drawback is that two words are only considered similar if they physically appear in the same document a certain number of times. For example consider the words "astronaut" and "cosmonaut". These words are certainly synonym but would never appear in the same document, at least not with a frequency recognized by a co-occurrence-based method. This kind of relationship can be found in a general-purpose hand-crafted thesaurus such as WordNet, and can also be found in a thesaurus constructed using head-modifier relations, since we can expect "astronaut" and "cosmonaut" to share the same context.

Although head-modifier-based corpus-derived thesauri do not display the above problems, words with similar heads or modifiers are not always good candidates for expansion [23]. For example, adjective referring to countries share similar heads (the Indonesian/Japanese capital, the Indonesian/Japanese government, etc), but adding "Japanese" to a query that contains "Indonesian" will rarely produce good results. Note that there are many words that distinguish Japanese and Indonesian in terms of intra-sentential co-occurrence and general-purpose thesaurus.

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 thesaurus: a hand-crafted general-purpose thesaurus (WordNet), an automatically constructed thesaurus based on a document co-occurrence relations (co-occurrence based thesaurus) and an automatically constructed thesaurus based on head-modifier relations (head-modifier 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 [14]. 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 the noun taxonomy.

The similarity between words a and b can be defined as the shortest path from each sense of $w1$ to each sense of $w2$, as below [17] :

$$sim_{path}(w1, w2) = max[-log(\frac{N_p}{2D})]$$

where N_p is the number of nodes in path p from $w1$ to $w2$ and D is the maximum depth of the taxonomy.

Similarity also can be measured using the information content of the concepts that subsume words in the taxonomy, as below [17] :

$$sim_{IC}(w1, w2) = \max_{c \in S(c1,c2)} [-\log p(c)]$$

where $S(c1, c2)$ is the set of concepts that subsume both c1 and c2.

Concept probabilities are computed simply as the relative frequency derived from the document collection.

$$p(c) = \frac{freq(c)}{N}$$

where N is the total number of nouns observed, excluding those not subsumed by any WordNet class.

We sum up the path-based similarity and information-content-based similarity to serve as the final similarity.

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.

We use a variable-length window-size based on the multi-paragraph topic segmentation proposed by Hearst [10, 8, 9]. The main algorithm has three main parts :

- tokenization
  The text is subdivided into pseudo-sentences of a predefined word size s.
- similarity determination
  k pseudo-sentences are grouped together into a block to be compared against an adjacent group of pseudo-sentences (adjacent block). Similarity values are computed for every pseudo-sentence gap number; that is, score is asigned to pseudo-sentence gap i corresponding to how similar the pseudo-sentences from pseudo-sentence $i - k$ through i are to the pseudo-sentence from $i + 1$ to $i + k + 1$. Similarity between blocks is calculated by a cosine measure: given two text blocks $b1$ and $b2$, each with k pseudo-sentences,

  $$sim(b1, b2) = \frac{\sum_t w_{t,b1} w_{t,b2}}{\sqrt{\sum_t w_{t,b1}^2 \sum_t w_{t,b2}^2}}$$

  where t ranges over all the terms that have been registered during the tokenization, and $w_{t,b1}$ is their frequency within the $b1$ block.

- Boundary identification
  Boundaries are determined by changes in the sequence of similarity scores. For a given pseudo-sentence gap i, the algorithm looks at the scores of the pseudo-sentence gaps to the left of i as long as their values are increasing. When the values to the left peak out, the difference between the score at the peak and the score at i is recorded. The same procedure is performed with the pseudo-sentence gaps to the right of i. Finally, the relative height of the peak to the right of i is added to the relative height of the peak to the left. These new scores, called $depth - scores$, correspond to how sharp a change occurs on both sides of the pseudo-sentence gap. After performing average smoothing, a boundary is determined by defining the cutoff as a function of the average and standard deviation of the $depth - scores$ for the text.

In this paper, we used the parameter as belows :

- Width of the pseudo-sentences (s) is 20
- Blocksize (k) is 6

After the topic-segments are determined, we use mutual information as a tool for computing similarity between words. Mutual information compares the probability of the co-occurence of words a and b with the independent probabilities of occurrence of a and b :

$$I(a, b) = log\frac{P(a, b)}{P(a)P(b)}$$

where the probabilities of $P(a)$ and $P(b)$ are estimated by counting the number of occurrences of a and b in topic-segments. The joint probability is estimated by counting the number of times that word a co-occurs with b.

3.3 Head-modifier-based thesaurus

In this method term relations are gathered on the basis of linguistic relations and not document co-occurrence statistics [11]. Words appearing in a similar grammatical context are assumed to be similar, and therefore classified into the same class.

First, all the documents are parsed using the Apple Pie Parser, which is a probabilistic chart parser developed by Satoshi Sekine [24].

Then the following syntactic structures are extracted.

- Subject-Verb
- Verb-Object
- Adjective-Noun
- Noun-Noun

Each noun has a set of verbs, adjectives, and nouns that it co-occurs with, and for each such relationship, a mutual information value is calculated.

- $I(v_i, sub, n_j) = log\frac{f_{sub}(n_j, v_i)/N_{sub}}{(f_{sub}(n_j)/N_{sub})(f(v_i)/N_{sub})}$
  where $f_{sub}(v_i, n_j)$ is the frequency of noun n_j occurring as the subject of verb v_i, $f_{sub}(n_j)$ is the frequency of the noun n_j occurring as subject of any verb, $f(v_i)$ is the frequency of the verb v_i, and N_{sub} is the number of subject-verb construction.
- $I(v_i, obj, n_j) = log\frac{f_{obj}(n_j, v_i)/N_{obj}}{(f_{obj}(n_j)/N_{obj})(f(v_i)/N_{obj})}$
  where $f_{obj}(v_i, n_j)$ is the frequency of noun n_j occurring as the object of verb v_i, $f_{obj}(n_j)$ is the frequency of the noun n_j occurring as object of any verb, $f(v_i)$ is the frequency of the verb v_i, and N_{sub} is the number of verb-object construction.
- $I(a_i, adj, n_j) = log\frac{f_{adj}(n_j, a_i)/N_{adj}}{(f_{adj}(n_j)/N_{adj})(f(a_i)/N_{adj})}$
  where $f(a_i, n_j)$ is the frequency of noun n_j occurring as the argument of adjective a_i, $f_{adj}(n_j)$ is the frequency of the noun n_j occurring as the argument of any adjective, $f(a_i)$ is the frequency of the adjective a_i, and N_{adj} is the number of adjective-noun construction.
- $I(n_i, noun, n_j) = log\frac{f_{noun}(n_j, n_i)/N_{noun}}{(f_{noun}(n_j)/N_{noun})(f(n_i)/N_{noun})}$
  where $f(a_i, n_j)$ is the frequency of noun n_j occurring as the argument of noun n_i, $f_{noun}(n_j)$ is the frequency of the noun n_j occurring as the argument of any noun, $f(n_i)$ is the frequency of the noun n_i, and N_{noun} is the number of noun-noun construction.

For generality, we can use the notation $|w, r, w'|$ to denote the frequency count of the dependency triples (w, r, w'), where w and w' are two words which bear the syntactic relation r as above. When w, r, or w' is *, the frequency counts af all dependency triples matching the rest of the pattern are summed up. Thereby, the amount of information of words $w1$ and $w2$, $I(w, r, w')$, can be computed as $\log \frac{|w,r,w'| \times |*,r,*|}{|w,r,*| \times |*,r,w'|}$

Let $T(w)$ is the set of pairs (r, w') such that $I(w, , r, w')$ is positive. The final similarity $sim(w,w_2)$ between two words w_1 and w_2 can be computed as follows :

$$\frac{\sum_{(r,w)\in T(w_1)\cap T(w_2)} (I(w_1, r, w) + I(w_2, r, w))}{\sum_{(r,w)\in T(w_1)} I(w_1, r, w) + \sum_{(r,w)\in T(w_2)} I(w_2, r, w)}$$

3.4 Combination and Term Expansion Method

A query q is represented by the vector $\overrightarrow{q} = (q_1, q_2, ..., q_n)$, where each q_i is the weight of each search term t_i contained in query q. We used SMART version 11.0 [19] to obtain the initial query weight using the formula ltc as below :

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

where tf_{ik} is the occurrrence frequency of term t_k in query q_i, 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.

Using the above weighting method, the weight of initial query terms lies between 0 and 1. On the other hand, the similarity in each type of thesaurus does not have a fixed range. Hence, we apply the following normalization strategy to each type of thesaurus to bring the similarity value into the range $[0, 1]$.

$$sim_{new} = \frac{sim_{old} - sim_{min}}{sim_{max} - sim_{min}}$$

The similarity value between two terms in the combined thesauri is defined as the average of their similarity value over all types of thesaurus.

A query q is represented in terms of a vector $\overrightarrow{q} = (w_1, w_2, ..., w_n)$, where n is the total number of terms in the collection and 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 [16]:

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

where $sim(t_i, t_j)$ is defined as the average of the similarities of the three types of thesaurus 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 $w_{ex}(q, t_j)$ of an expansion term t_j with respect to a query q is defined as the following function of $sim_{qt}(q, t_j)$:

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

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

$$\overrightarrow{q}_e = (w'_1, w'_2, ..., w'_n),$$

to the original query vector. w'_j is equal to $w_{ex}(q, t_j)$ if $w_{ex}(q, t_j)$ exceeds some weight threshold, 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 [28]. 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 Figure 1 shows an example of a topic.

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

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 [19] 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. We use *lnc* weighting method for document collection as follows :

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

We further use the *ltc* weighting method for the initial query 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 occurrrence frequency of term t_k in query q_i (for query term weighting) or in document d_i (for document term 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. We use 0.1 as weight threshold (decided experimentally) and fixed for all queries.

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 thesaurus types. For each method we give the percentage of improvement over the baseline method in parentheses. We demonstrate that the performance using the combined thesauri for query expansion is better than SMART and also than expansion using just one type of thesaurus.

We investigated to what extent each thesaurus contributes to the provision of expansion terms. Table 4 summarizes the percentage of expansion terms that are added using each thesaurus and different combinations of thesauri. We can see that each thesaurus contributes almost the same in providing expansion terms.

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

- Dice Coefficient [13]
  Suppose two words a and b occur in df_a and df_b windows respectively, and co-occur in df_c windows, then the similarity of a and b is calculated as follows:
  $$\frac{2 \times df_c}{df_a + df_b}.$$
- Tanimoto coefficient
  Suppose two words a and b occur in df_a and df_b windows respectively, and co-occur in df_c windows, then the similarity of a and b is calculated as follows:
  $$\frac{df_c}{df_a + df_b - df_c}$$

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

Title:
clothing sweatshops
Description:

Identify documents that discuss clothing sweatshops.

Narrative:
A relevant document must identify the country, the working conditions, salary, and type of clothing or shoes being produced. Relevant documents may also include the name of the business or company or the type of manufacturing, such as: "designer label".

Figure 1: Topics 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	Head-Mod only	Cooccur only	WordNet+ Head-Mod	WordNet+ Cooccur	Head-Mod+ Cooccur	Combined method
Title	0.1175	0.1299 (+10.6%)	0.1505 (+28.1%)	0.1637 (+39.3%)	0.1611 (+37.1%)	0.1698 (+44.5%	0.1859 (+58.2%)	0.2337 (+98.9%)
Description	0.1428	0.1525 (+6.8%)	0.1705 (+19.4%)	0.1950 (+33.4%)	0.1832 (+28.3%)	0.1973 (+38.2%)	0.2315 (+62.1%)	0.2689 (+88.3%)
All	0.1976	0.2018 (+2.1%)	0.2249 (+13.8%)	0.2395 (+21.2%)	0.2276 (+15.2%)	0.2423 (+22.6%)	0.2565 (+29.8%)	0.2751 (+39.2%)

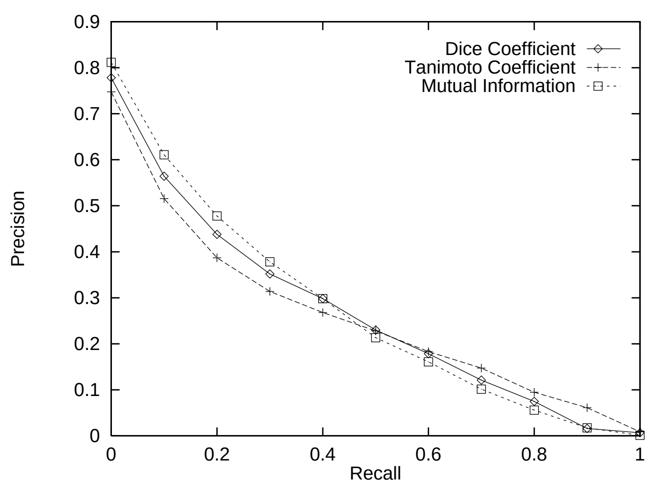

Figure 2: 11-point precision for different similarity measures

Table 4: Contribution of each thesaurus for expansion terms

Thesaurus	Title (%)	Description (%)	All (%)
WordNet	6	7	5
Co-occurrence	12	9	20
Head-Modifier	10	8	14
WordNet and Co-occurrence	14	13	9
WordNet and Head-Modifier	13	13	7
Co-occurrence and Head-Modifier	17	28	32
All	28	22	13

Table 5: Comparison of different similarity measures for different topic sections

Topic section	Base	Dice Coef.	Mutual Information	Tanimoto Coef.
Title	0.1175	0.2213 (+88.3%)	0.2337 (+98.9%)	0.2011 (+71.7%)
Description	0.1428	0.2590 (+81.4%)	0.2689 (+88.3%)	0.2492 (+75.3%)
All	0.1976	0.2573 (+30.2%)	0.2751 (+39.2%)	0.2654 (+34.5%)

Figure 2 shows a recall-precision graph using the description section of the topics, and Table 5 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 Discussion

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

- broadening thesaurus coverage by combining different types of thesauri
- weighting expansion terms to eliminate misleading expansion term

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 thesaurus.

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 repository for money and a pile of earth on the edge of a river. 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 relationships with other terms in the original query, resulting in a low weight.

6 Conclusions

We have proposed the use of different types of thesaurus as evidence 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 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 thesaurus. This is actually a kind of word sense disambiguation.

Experiments have shown that combined use of thesauri gives better retrieval results than using just one type of thesaurus.

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

7 Acknowledgements

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

References

[1] C. Buckley and G. Salton. The effect of adding relevance information in a relevance feedback environment. In *Proceedings of the 17th ACM SIGIR Conference*, pp. 292–300, 1994.

[2] H. Chen, B. Schatz, T. Yim, and D. Fye. Automatic thesaurus generation for an electronic community system. *Journal of American Society for Information Science*, Vol. 46, No. 3, pp. 175–193, 1995.

[3] Carolyn J. Crouch. An aproach to the automatic construction of global thesauri. *Information Processing and Management*, Vol. 26, No. 5, pp. 629–640, 1990.

[4] F.C. Ekmekcioglu. Effectiveness of query expansion in ranked-output document retrieval systems. *Journal of Information Science*, Vol. 18, pp. 139–147, 1992.

[5] Edward A. Fox. Lexical relations enhancing effectiveness of information retrieval systems. *SIGIR Forum*, Vol. 15, No. 3, pp. 6–36, 1980.

[6] Gregory. Grafenstette. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publisher, 1994.

[7] G. Grefenstette. Use of syntactic context to produce term association lists for text retrieval. In *Proceedings of the 15th ACM SIGIR Conference*, pp. 89–97, 1992.

[8] Marti A. Hearst. Multi-paragraph segmentation of expository text. In *Proceedings of 32th Annual Meeting of the ACL*, pp. 9–16, 1994.

[9] Marti A. Hearst. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, Vol. 23, No. 1, pp. 33–64, 1997.

[10] Marti A Hearst and Christian Plaunt. Subtopic structuring for full-length document access. In *Proceedings of 16th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pp. 59–68, 1993.

[11] D. Hindle. Noun classification from predicate-argument structures. In *Proceedings of 28th Annual Meeting of the ACL*, pp. 268–275, 1990.

[12] Y. Jing and B. Croft. An association thesaurus for information retrieval. In *Proceedings of RIAO*, pp. 146–160, 1994.

[13] R. Mandala, T. Tokunaga, H. Tanaka, A. Okumura, and K. Satoh. Adhoc retrieval experiments using WordNet and automatically constructed thesauri. In *Proceedings of the Seventh Text REtrieval Conference (TREC-7)*, 1999. NIST Special Publication.

[14] G.A Miller. Wordnet: An on-line lexical database. *Special Issue of the International Journal of Lexicography*, Vol. 3, No. 4, 1990.

[15] J.H. Peat and P. Willett. The limitations of term co-occurrence data for query expansion in document retrieval systems. *Journal of American Society for Information Science*, Vol. 42, No. 5, pp. 378–383, 1991.

[16] Qiu and H.P. Frei. Concept based query expansion. In *Proceedings of the 16th ACM SIGIR Conference*, pp. 160–169, 1993.

[17] Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI-95)*, pp. 448–453, 1995.

[18] G Ruge. Experiments on linguistically-based term associations. *Information Processing and Management*, Vol. 28, No. 3, pp. 317–332, 1992.

[19] G. Salton. *The SMART Retrieval System Experiments in Automatic Document Processing*. Prentice-Hall, 1971.

[20] G. Salton and C. Buckley. Improving retrieval performance by relevance feedback. *Journal of American Society for Information Science*, Vol. 41, No. 4, pp. 288–297, 1990.

[21] G. Salton and M McGill. *An Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.

[22] H. Schutze and J.O. Pederson. A cooccurrence-based thesaurus and two applications to information retrieval. In *Proceedings of RIAO Conference*, pp. 266–274, 1994.

[23] H. Schutze and J.O Pederson. A cooccurrence-based thesaurus and two applications to information retrieval. *Information Processing and Management*, Vol. 33, No. 3, pp. 307–318, 1997.

[24] S. Sekine and R. Grishman. A corpus-based probabilistic grammar with only two non-terminals. In *Proceedings of the International Workshop on Parsing Technologies*, 1995.

[25] A.F. Smeaton and C. Berrut. Thresholding postings lists, query expansion by word-word distances and POS tagging of Spanish text. In *Proceedings of The Fourth Text Retrieval Conference*, 1996.

[26] E.M. Voorhees. Using WordNet to disambiguate word senses for text retrieval. In *Proceedings of the 16th ACM-SIGIR Conference*, pp. 171–180, 1993.

[27] E.M. Voorhees. Query expansion using lexical-semantic relations. In *Proceedings of the 17th ACM-SIGIR Conference*, pp. 61–69, 1994.

[28] E.M. Voorhees and D. Harman. Overview of the seventh text retrieval conference (TREC-7). In *Proceedings of the Seventh Text REtrieval Conference*, 1999. NIST Special Publication.