The Exploration and Analysis of Using Multiple Thesaurus Types for Query Expansion in Information Retrieval

Rila Mandala,† Takenobu Tokunaga† and Hozumi Tanaka†

This paper proposes the use of multiple thesaurus types for query expansion in information retrieval. Hand-crafted thesaurus, corpus-based co-occurrence-based thesaurus and syntactic-relation-based thesaurus are combined and used as a tool for query expansion. A simple word sense disambiguation is performed to avoid misleading expansion terms. Experiments using TREC-7 collection proved that this method could improve the information retrieval performance significantly. Failure analysis was done on the cases in which the proposed method fail to improve the retrieval effectiveness. We found that queries containing negative statements and multiple aspects might cause problems in the proposed method.

KeyWords: *multiple thesaurus types, query expansion, information retrieval*

1 Introduction

The task of information retrieval system is to extract relevant documents from a large collection of documents in response to user queries (Salton and McGill 1983). Most modern information retrieval systems do not output a set of documents for a query. Instead, they output a list of documents ranked in descending order of relevance to the query (Baeza-Yates and Ribeiro-Neto 1999). In consequence, the task of modern information retrieval system can be re-stated as to push the relevant documents to the top of the retrieved documents rank.

Although information can be presented in diverse form such as tabular numerical data, graphical displays, photographic images, human speech, and so on, the term *information retrieval* as used in this paper shall refer specifically to the retrieval of textual information.

The fundamental problems in information retrieval is that there are many ways to express the same concept in natural language (Blair and Maron 1985; Grossman and Frieder 1998). User in different contexts, or with different information needs or knowledge often describe the same information using different terms. In consequence, relevant document which do not contain the exact terms as the query will be put in low rank.

In this paper, we address the word mismatch problem through automatic query expansion (Ekmekcioglu 1992). The query is expanded by using terms which have related meaning to

† Department of Computer Science, Graduate School of Information Science and Engineering, Tokyo Institute of Technology

those in the query. The expansion terms can be taken from thesauri (Aitchison and Gilchrist 1987; Paice 1991; Kristensen 1993). Roughly, there are two types of thesauri, i.e., hand-crafted thesauri and corpus-based automatically constructed thesauri. Hand-crafted thesauri describe the synonymous relationship between words, though many thesauri use finer grained relation such as broader terms, narrower terms, and so on. Some of the hand-crafted thesauri are for specific domains while others are for general purpose. The relation in hand-crafted thesauri can be used for query expansion. Query expansion using specific domain thesauri has been reported yielding a very good results (Fox 1980; Chen, Schatz, Yim, and Fye 1995). Currently, the number of existing domain specific thesauri can be counted by finger, while the number of domain in the world is very large. Unfortunately building such thesauri manually requires a lot of human labor from linguists or domain experts and spending very much time. In contrast, the use of general-purpose thesauri for query expansion has been reported fail to improve the information retrieval performance by several researchers (Richardson and Smeaton 1994, 1995; Voorhees 1994, 1988; Smeaton and Berrut 1996; Stairmand 1997).

Automatic thesaurus construction is an extensive studied area in computational linguistics (Charniak 1993; Church and Hanks 1989; Hindle 1990; Lin 1998). The original motivation behind the automatic thesaurus construction is to find an economic alternative to hand-crafted thesaurus. Broadly, there are two methods to construct thesaurus automatically. The first one is based on the similarities between words on the basis of co-occurrence data in each document (Qiu and Frei 1993; Schutze and Pederson 1994, 1997; Crouch 1990; Crouch and Yang 1992), and the other one is based on the co-occurrence of some syntactic relations such as predicate-argument in the whole documents (Jing and Croft 1994; Ruge 1992; Grefenstette 1992; Grafenstette 1994; Hindle 1990).

Many researcher found some slight improvement using the co-occurrence-based thesaurus (Qiu and Frei 1993; Schutze and Pederson 1997), and some mixed results using the syntactic-relation-based thesaurus (Jing and Croft 1994; Grafenstette 1994).

Previously, we conducted an analysis of the different types of thesauri described above, and found that each type of thesaurus has different advantages and disadvantages (Rila Mandala, Tokunaga, Tanaka, Okumura, and Satoh 1999d; Rila Mandala, Tokunaga, and Tanaka 1999c, 1999a, 1999b) which can be summarized as follows :

- Hand-crafted thesaurus
  - can capture general term relation.
  - can not capture domain-specific relation.
- Co-occurrence-based thesaurus

  - can capture domain-specific relation.
  - can not capture the relation between terms which do not co-occur in the same document or window.
- Syntactic-relation-based thesaurus
  - can capture domain-specific relation.
  - can capture the relation between terms even though they do not co-occur in the same document.
  - words with similar heads or modifiers are not always good candidates for expansion

In this paper we explore and analyze a method to combine the three types of thesauri (hand-crafted, co-occurrence-based, and syntactic-relation-based thesaurus) for the purpose of query expansion. In the next section we desribe the detail method of combining thesauri, and in Section 3 we give some experimental results using a large TREC-7 collection and several small information retrieval test collections. We discuss why our method work in Section 4 and also perform failure analysis in Section 5. We tried to combine our method with pseudo-relevance-feedback along with experimental results in Section 6. Finally, in Section 7 we give conclusions and future work.

2 Method

In this section, we first describe our method to construct each type of thesaurus utilized in this research, and then describe our attemp to minimize the misleading expansion terms by using term weighting method based on these thesauri.

2.1 WordNet

WordNet is a machine-readable hand-crafted thesaurus (Miller 1990). Word forms in WordNet are represented in their familiar orthograpy and word meanings are represented by synonym sets (synset) (Fellbaum 1998). A synonym set represents a concept and comprises all those terms which can be used to express the concept. In other word a synset is a list of synonymous word forms that are interchangeable in some context.

The similarity between words w_1 and w_2 can be defined as the shortest path from each sense of w_1 to each sense of w_2, as below (Leacock and Chodorow 1988) :

$$sim_{path}(w_1, w_2) = max[-\log(\frac{N_p}{2D})]$$

where N_p is the number of nodes in path p from w_1 to w_2 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 (Resnik 1995) :

$$sim_{IC}(w_1, w_2) = \max_{c \in S(c_1, c_2)} [-\log p(c)]$$

where $S(c_1, c_2)$ is the set of concepts that subsume both c_1 and c_2.

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.

2.2 Co-occurrence-based thesaurus

Co-occurrence-based thesaurus utilize the number of occurrence or co-occurrence of words within a document or within a window as a source of information to build thesaurus. We use textual windows based on TextTiling algorithm (Hearst 1994, 1997) to calculate the mutual information between a pair of words. TextTiling is a paragraph-level model of discourse structure based on the notion of subtopic shift and an algorithm for subdividing expository text into multi-paragraph passages or subtopic segments. This algorithm makes use of patterns of lexical co-occurrence and distribution. The algorithm has three parts: tokenization into terms and sentence-sized units, determination of a score for each sentence-sized unit, and detection of the subtopic boundaries, which are assumed to occur at the largest valleys in the graph that results from plotting sentence-unit against scores. We then employ an information theoretic definition of mutual information which compares the probability of observing two words together to that of observing each word independently in the passages defined by TextTiling. Words having high mutual information over a corpus are assumed semantically related.

2.3 Syntactic-relation-based Thesaurus

The basic premise of this method to build thesaurus is that words found in the same grammatical context tend to share semantic similarity. Syntactic analysis allows us to know what words modify other words, and to develop contexts from this information (Grafenstette 1994;

Ruge 1992; Hindle 1990).

To build such thesaurus, firstly, all the documents are parsed using the Apple Pie Parser (Sekine and Grishman 1995). This parser is a bottom-up probabilistic chart parser which finds the parse tree with the best score by way of the best-first search algorithm. Its grammar is a semi-context sensitive grammar with two non-terminals and was automatically extracted from Penn Tree Bank syntactically tagged corpus developed at the University of Pennsylvania. The parser generates a syntactic tree in the manner of a Penn Tree Bank bracketing. The accuracy of this parser is reported as parseval recall 77.45 % and parseval precision 75.58 %.

Using the above parser, we extracted subject-verb, verb-object, adjective-noun, and noun-noun relations, so that 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.

- $I_{sub}(v_i, n_j) = \log \frac{f_{sub}(n_j,v_i)/N_{sub}}{(f_{sub}(n_j)/N_{sub})(f(v_i)/N_{sub})}$
  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 relations.
- $I_{obj}(v_i, n_j) = \log \frac{f_{obj}(n_j,v_i)/N_{obj}}{(f_{obj}(n_j)/N_{obj})(f(v_i)/N_{obj})}$
  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_{obj} is the number of verb-object relations.
- $I_{adj}(a_i, n_j) = \log \frac{f_{adj}(n_j,a_i)/N_{adj}}{(f_{adj}(n_j)/N_{adj})(f(a_i)/N_{adj})}$ 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 relations.
- $I_{noun}(n_i, n_j) = \log \frac{f_{noun}(n_j,n_i)/N_{noun}}{(f_{noun}(n_j)/N_{noun})(f(n_i)/N_{noun})}$ where $f(n_i, n_j)$ 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 relations.

The similarity between two words w_1 and w_2 can be computed as follows :

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

where r is the syntactic relation type, and w is

- a verb, if r is the subject-verb or object-verb relation.
- an adjective, if r is the adjective-noun relation.

- a noun, if r is the noun-noun relation.

and $T(w)$ is the set of pairs (r, w') such that $I_r(w, w')$ is positive.

2.4 Combination and Term Expansion Method

A query q is represented by the vector $\overrightarrow{\mathrm{q}} = (w_1, w_2, ..., w_n)$, where each w_i is the weight of each search term t_i contained in query q. We used SMART version 11.0 (Salton 1971) to obtain the initial query weight using the formula *ltc* as belows :

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

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 $[0, 1]$.

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

Although there are many combination methods that can be tried, we just define the similarity value between two terms in the combined thesauri as the average of their similarity value over all types of thesaurus because we do not want to introduce additional parameters here which depend on queries nature.

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

$$simqt(q, t_j) = \sum_{t_i \in q} w_i * sim(t_i, t_j)$$

where the value of $sim(t_i, t_j)$ is taken from the combined thesauri as described above.

With respect to the query q, all the terms in the collection can now be ranked according to their *simqt*. Expansion terms are terms t_j with high $simqt(q, t_j)$.

The $weight(q, t_j)$ of an expansion term t_j is defined as a function of $simqt(q, t_j)$:

$$weight(q, t_j) = \frac{simqt(q, t_j)}{\sum_{t_i \in q} w_i}$$

where $0 \leq weight(q, t_j) \leq 1$.

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 the similarities between the term t_j and all the query terms. The weight of the original query terms are the weighting factors of those similarities.

Therefore the query q is expanded by adding the following query :

$$\overrightarrow{\mathrm{q_e}} = (a_1, a_2, ..., a_r)$$

where a_j is equal to $weight(q, t_j)$ if t_j belongs to the top r ranked terms. Otherwise a_j is equal to 0.

The resulting expanded query is :

$$\overrightarrow{\mathrm{q}}_{expanded} = \overrightarrow{\mathrm{q}} \circ \overrightarrow{\mathrm{q_e}}$$

where the $\circ$ is defined as the concatenation operator.

The method above can accommodate polysemy, because an expansion term which is taken from a different sense to the original query term is given a very low weight.

3 Experimental Results

3.1 Test Collection

As a main test collection we use TREC-7 collection (Voorhees and Harman 1999). TREC (Text REtrieval Conference) is an DARPA (Defense Advanced Research Project Agency) and NIST (National Institute of Standards and Technology) co-sponsored effort that brings together information retrieval researchers from around the world to discuss and compare the performance of their systems, and to develop a large test collection for information retrieval system. The seventh in this series of annual conferences, TREC-7, attracted 56 different participants from academic institutions, government organizations, and commercial organizations (Voorhees and Harman 1999). With such a large participation of various information retrieval researchers, a large and varied collections of full-text documents, a large number of user queries, and a superior set of independent relevance judgements, TREC collections have rightfully become the standard test collections for current information retrieval research.

The common information retrieval task of ranking documents for a new query is called the *adhoc* task in the TREC framework. The TREC data comes on CD-ROMs, called the TREC disks. The disks are numbered, and a combination of several disk can be used to form a text collection for experimentation.

The TREC-7 test collection consists of 50 topics (queries) and 528,155 documents from

Table 1 TREC-7 Document statistics

Source	Size (Mb)	Number of documents	Average number of terms/article
Disk 4			
The Financial Times, 1991-1994 (FT)	564	210,158	412.7
Federal Register, 1994 (FR94)	395	55,630	644.7
Disk 5			
Foreign Broadcast Information Services (FBIS)	470	130,471	543.6
the LA Times	475	131,896	526.5

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 TREC-7 document collection, Table 2 shows statistics of the topics, and Figure 1 shows an example of a topic, and Figure 2 shows its expansion terms produced by our method.

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

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".

Fig. 1 Topics Example

wage	labor	sewing	low	minimum	payment
earning	workshop	workplace	shop	welfare	county
circumstance	overtime	child	entrepreneur	employment	manufacture
immigrant	industry	bussiness	company	violation	remuneration
apparel	vesture	wear	footwear	footgear	enterprise
commercialism	machine	status	plant	raise	production
calcitonin					

Fig. 2 Expansion terms example

It is well known that many information retrieval techniques are sensitive to factors such as query length, document length, and so forth. For example, one technique which works very well for long queries may not work well for short queries. To ensure that our techniques and conclusions are general, we use different-length query in TREC-7 collection.

Beside the large and the newer TREC-7 test collection described before, we also use some previous small test collections (Fox 1990), because although most real world collections are large, some can be quite small. These small collections have been widely used in the experiments by many information retrieval researchers before TREC. These old test collections have always been built to serve some purpose. For example, the Cranfield collection was originally built to test different types of manual indexing, the MEDLINE collection was built in an early attempt to compare the operational Boolean MEDLARS system with the experimental ranking used in SMART, and the CACM and CISI collections were built to investigate the use of an extended vector space model that included bibliographic data. Most of the old test collections are very domain specific and contain only the abstract.

In Table 3 and 4 we describe the statistics and the domain of the old collection, respectively.

3.2 Evaluation method

Recall and precision are two widely used metrics to measure the retrieval effectiveness of an information retrieval system. Recall is the fraction of the relevant documents which has been retrieved, i.e.

$$recall = \frac{number\ of\ relevant\ documents\ retrieved}{number\ of\ relevant\ documents\ in\ collection}.$$

Table 3 Small collection statistics

Collection	Number of Documents	Average Terms/Docs	Number of Query	Average Terms/query	Average Relevant/query
Cranfield	1398	53.1	225	9.2	7.2
ADI	82	27.1	35	14.6	9.5
MEDLARS	1033	51.6	30	10.1	23.2
CACM	3204	24.5	64	10.8	15.3
CISI	1460	46.5	112	28.3	49.8
NPL	11429	20.0	100	7.2	22.4
INSPEC	12684	32.5	84	15.6	33.0

Table 4 The domain of the small collections

Collection	Domain
Cranfield	Aeronautics
ADI	Information Science
MEDLINE	Medical Science
CACM	Computer Science
CISI	Computer and Information Science
NPL	Electrical Engineering
INSPEC	Electrical Engineering

Precision is the fraction of the retrieved document, i.e.

$$precision = \frac{number\ of\ relevant\ documents\ retrieved}{total\ number\ of\ documents\ retrieved}.$$

However, precision and recall are set-based measures. That is, they evaluate the quality of an unordered set of retrieved documents. To evaluate ranked lists, precision can be plotted against recall after each retrieved document. To facilitate comparing performance over a set of topics, each with a different number of relevant documents, individual topic precision values are interpolated to a set of standard recall levels (0 to 1 in increments of 0.1). The particular rule used to interpolate precision at standard recall level i is to use the maximum precision obtained for the topic for any actual recall level greater than or equal to i. Note that while precision is not defined at a recall 0.0, this interpolation rule does define an interpolated value for recall level 0.0. For example assume a document collection has 20 documents, four of which are relevant to topic t in which they are retrieved at ranks 1, 2, 4, 15. The exact recall points are 0.25, 0.5, 0.75, and 1.0. Using the interpolation rule, the interpolated precision for all standard recall levels 0.0, 0.1, 0.2, 0.3, 0.4, and 0.5 is 1, the interpolated precision for recall levels 0.6 and 0.7 is 0.75, and the interpolated precision for recall levels 0.8, 0.9, and 1.0 is 0.27.

3.3 Results

Table 5 shows the average of 11-point interpolated precision using various section of topics in TREC-7 collection, and Table 6 shows the average of 11-point interpolated precision in several small collections. We can see that our method give a consistent and significant improvement compared with the baseline and using only one type of thesaurus.

Table 5 Experiment results using TREC-7 Collection

Topic Type	Base	Expanded with						
		WordNet only	Syntactic only	Cooccur only	WordNet+ Syntactic	WordNet+ Cooccur	Syntactic+ Cooccur	Combined method
Title	0.1452	0.1541 (+6.1%)	0.1802 (+24.1%)	0.1905 (+31.2%)	0.1877 (+29.3%)	0.2063 (+42.1%)	0.2197 (+51.3%)	0.2659 (+83.1 %)
Description	0.1696	0.1777 (+4.8%)	0.1974 (+16.4%)	0.2144 (+26.4%)	0.2057 (+21.3%)	0.2173 (+28.1%)	0.2337 (+37.8%)	0.2722 (+60.5 %)
All	0.2189	0.2235 (+2.1%)	0.2447 (+11.8%)	0.2566 (+17.2%)	0.2563 (+17.1%)	0.2611 (+19.3%)	0.2679 (+22.4%)	0.2872 (+31.2 %)

Table 6 Experiment results using small collection

Coll	Base	Expanded with						
		WordNet only	Syntactic only	Cooccur only	WordNet+ Syntactic	WordNet+ Cooccur	Syntactic+ Cooccur	Combined method
ADI	0.4653	0.4751 (+2.1%)	0.5039 (+8.3%)	0.5146 (+10.6%)	0.5263 (+13.1%)	0.5486 (+17.9%)	0.5895 (+26.7%)	0.6570 (+41.2%)
CACM	0.3558	0.3718 (+4.5%)	0.3853 (+8.3%)	0.4433 (+24.6%)	0.4109 (+15.5%)	0.4490 (+26.2%)	0.4796 (+34.8%)	0.5497 (+54.5%)
INSPEC	0.3119	0.3234 (+3.7%)	0.3378 (+8.3%)	0.3755 (+20.4%)	0.3465 (+11.1%)	0.4002 (+28.3%)	0.4420 (+41.7%)	0.5056 (+62.1 %)
CISI	0.2536	0.2719 (+7.2%)	0.2800 (+10.4%)	0.3261 (+28.6%)	0.3076 (+21.3%)	0.3606 (+42.2%)	0.4009 (+58.1%)	0.4395 (+73.3 %)
CRAN	0.4594	0.4700 (+2.3%)	0.4916 (+7.0%)	0.5435 (+18.3%)	0.5012 (+9.1%)	0.5706 (+24.2%)	0.5931 (+29.1%)	0.6528 (+42.1 %)
MEDLINE	0.5614	0.5681 (+1.2%)	0.6013 (+7.1%)	0.6372 (+13.5%)	0.6114 (+8.9%)	0.6580 (+17.2%)	0.6860 (+22.2%)	0.7551 (+34.5%)
NPL	0.2700	0.2840 (+5.2%)	0.2946 (+9.1%)	0.3307 (+22.5%)	0.3038 (+12.5%)	0.3502 (+29.7%)	0.3796 (+40.6%)	0.4469 (+65.5%)

4 Discussion

The important points of our method are :

- the coverage of WordNet is broadened
- weighting method.

The three types of thesauri we used have different characteristics. Automatically constructed thesauri add not only new terms but also new relationships not found in WordNet. If two terms often co-occur together in a document then those two terms are likely bear some relationship. Why not only use the automatically constructed thesauri ? The answer to this is that some relationships may be missing in the automatically constructed thesauri (Grafenstette 1994).

For example, consider the words *tumor* and *tumour*. These words certainly share the same context, but would never appear in the same document, at least not with a frequency recognized by a co-occurrence-based method. In general, different words used to describe similar concepts may never be used in the same document, and are thus missed by the co-occurrence methods. However their relationship may be found in the WordNet thesaurus.

The second point is our weighting method. As already mentioned before, most attempts at automatically expanding queries by means of WordNet have failed to improve retrieval effectiveness. The opposite has often been true: expanded queries were less effective than the original queries. Beside the "incomplete" nature of WordNet, we believe that a further problem, the weighting of expansion terms, has not been solved. All weighting methods described in the past researches of query expansion using WordNet have been based on "trial and error" or ad-hoc methods. That is, they have no underlying justification.

The advantages of our weighting method are:

- the weight of each expansion term considers the similarity of that term with all terms in the original query, rather than to just one or some query terms.
- the weight of the expansion term accommodates the polysemous word problem.

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 the financial institution and the river edge senses. 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 articles 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 concerned with finance and not rivers. Thus rivers would only have a relationship with the *bank* term and there would be no relationships with other terms in the original query, resulting in a low weight. Since our weighting method depends on both query in its entirety and similarity in the three thesauri, the wrong sense expansion terms are given very low weight.

5 Failure Analysis

Although our method as a whole gives a very significant improvement, it still further can be improved. Of the 50 queries of TREC-7 collection, our method improves the performance of 43 queries and degrade the performance of 7 queries compared with the baseline. We investigated manually why our method degrade the performance of several queries.

5.1 Negation statements in the query

We found that most of the queries hurted by our method contains the negation statements. Through our method, all the terms in the negation statements are also considered for query expansion which is degrading the retrieval performance for that query. Figure 3 shows two examples of query which contain negation statements.

Table 7 shows the results of eliminating the negation statements from the queries manually for each query containing negation statements. As that table shown, eliminating the negation statements improves the retrieval effectiveness. It is to be investigated further how we could identify the negation statements automatically.

Table 7 The results of negation statements elimination

Query Number	SMART	Expansion without Negation Elimination	Expansion with Negation Elimination
2	0.3643	0.3381 (- 7.19%)	0.3811 (+ 4.61%)
5	0.3112	0.2804 (- 9.90%)	0.3314 (+ 6.49%)
13	0.1621	0.1567 (- 3.33%)	0.1823 (+12.46%)
17	0.2310	0.2235 (- 3.25%)	0.2441 (+ 5.67%)
42	0.2732	0.2569 (- 5.97%)	0.2942 (+ 7.69%)
43	0.3031	0.2834 (- 6.50%)	0.3321 (+ 9.57%)

5.2 Multiple aspects of query

An examination of the top-ranked non-relevant documents for various queries shows that a commonly occurring cause of non-relevance among such documents is inadequate query

Title:

British Chunnel impact

Description:

What impact has the Chunnel had on the British economy and/or the life style of the British?

Narrative:

Documents discussing the following issues are relevant:
- projected and actual impact on the life styles of the British
- Long term changes to economic policy and relations
- major changes to other transportation systems linked with the Continent

Documents discussing the following issues are not relevant:
- expense and construction schedule
- routine marketing ploys by other channel crossers (i.e., schedule changes, price drops, etc.)

Title:

Ocean remote sensing

Description:

Identify documents discussing the development and application of spaceborne ocean remote sensing.

Narrative:

Documents discussing the development and application of spaceborne ocean remote sensing in oceanography, seabed prospecting and mining, or any marine-science activity are relevant. Documents that discuss the application of satellite remote sensing in geography, agriculture, forestry, mining and mineral prospecting or any land-bound science are not relevant, nor are references to international marketing or promotional advertizing of any remote-sensing technology. Synthetic aperture radar (SAR) employed in ocean remote sensing is relevant.

Fig. 3 Two examples of query containing negation statements

coverage, i.e., the query consists of multiple aspects, only some of which are covered in these documents. For example, a query of the TREC collection asks : *Identify documents discussing the use of estrogen by postmenopausal women in Britain.* Several top-ranked non-relevant documents contain information about the use of hormone by postmenopausal women but not in Britain. If we look at the expansion terms produced by our method as shown in Figure 4 we could see that many expansion terms have relationship with all query terms except Britain. This is because all query terms but Britain have relationship between each other and these terms have a high original term weight. On the contrary, Britain does not have relationship with other query terms and Britain have a low original term weight in almost all documents in collection. Consequently, the term related to Britain are given a low weight by our method.

estradiol	female	hormone	disease	therapy	menopausal
chemical	progesterone	menstruation	vaginal	progestin	obstetrics
gynecology	replacement	endometrial	cancer	breast	ovary
treatment	old	tamoxifen	symptom	synthetic	drug
hot	flash	osteoporosis	cholesterol	recepter	risk
calcium	bones	mineralization	medical	physiologist	diagnostic
calcitonin					

Fig. 4 Expansion terms

To investigate the relatedness or independence of query words, we examine their co-occurrence patterns in 1000 documents initially retrieved for a query. If two words have the same aspect, then they often occur together in many of these documents. If one of the words appears in a document, the chance of the other occurring within the same document is likely to be relatively high. On the other hand, if two words bear independent concepts, the occurrences of the words are not strongly related.

Based on this observation, we re-rank the top-1000 retrieved documents, by re-computing the similarity between a query $\overrightarrow{q} = \{t_1, t_2, .., t_m\}$ (terms are ordered by decreasing of their inverse document frequency) and document D as belows (Mitra, Singhal, and Buckley 1998) :

$$Sim_{new}(D) = idf(t_1) + \sum_{i=2}^{m} idf(t_i) \ \times \ min_{j=1}^{i-1}(1 - P(t_i|t_j)),$$

where *idf* is the inverse of document frequency in the top-1000 initially retrieved documents, m is the number of terms in query that appear in document D, and $P(t_i|t_j)$ is estimated based

on word occurrences in document collection and is given by :

$$\frac{\#\ documents\ containing\ words\ t_i\ and\ t_j}{\#\ documents\ containing\ word\ t_j}.$$

For example, in the query stated above, the terms *estrogen, postmenopausal,* and *women* are strongly related to each other. If the term *postmenopausal* occurs in a document, the probability of word *women* occurring in the same document is high. Accordingly, the contribution of word *women* to Sim_{new} is reduced in this case. On the other hand, terms *postmenopausal* and *Britain* correspond to two independent aspects of the query and the occurrences of these two terms are relatively uncorrelated. Therefore, if a document contains these two terms, the contribution of *Britain* is higher and it counts as an important new matching term since its occurrence is not well predicted by other matching term (*postmenopausal*). This technique can improve the average of 11-point interpolated precision of TREC-7 collection for about 3.3% as shown in Table 8.

We also investigated another method to overcome this problem in which we built a Boolean expression for all query manually. Terms in the same aspect of query are placed in *or* relation, and terms in different aspect are placed in *and* relation (Hearst 1996). Documents that satisfy the constraint contain at least one word from each aspect of the query. For example, for the query stated before (*Identify documents discussing the use of estrogen by postmenopausal women in Britain*), we construct boolean expression as follows :

```
estrogen and (postmenopausal or woman) and britain.
```

Using this method, we again re-rank the top 1000 documents initially retrieved. Documents that match more words in different aspect of query are ranked ahead of documents that match less words. Ties are resolved by referring to the original document weight. Using this method we can improve the average of 11-point interpolated precision of TREC-7 collection for about 11.3%, as shown in Table 8.

This correlation and boolean reranking methods degrade some queries performance, because in those queries these methods overweight several query terms.

It is to be further investigated how we could design the appropriate method to overcome this problem.

6 Combining with relevance feedback

In this section, we describe the combination of our method with pseudo-relevance feedback (Buckley and Salton 1994, 1995; Salton and Buckley 1990). Pseudo-relevance feedback

Table 8 The effect of re-ranking the top-1000 ranked initially retrieved using co-occurrence method and boolean filter method

Query Number	Without Re-ranking	Re-ranking correlation	%improvement	Reranking Boolean	%improvement
1	0.5153	0.5666	+9.96	0.7724	+49.89
2	0.3794	0.1952	-48.55	0.4740	+24.93
3	0.3230	0.2719	-15.82	0.3237	+0.22
4	0.2280	0.2731	+19.78	0.2355	+3.29
5	0.3213	0.2457	-23.53	0.2931	-8.78
6	0.0646	0.0495	-23.37	0.0655	+1.39
7	0.3878	0.5632	+45.23	0.3607	-6.99
8	0.2983	0.4270	+43.14	0.3049	+2.21
9	0.0422	0.0612	+45.02	0.0254	-39.81
10	0.2196	0.3223	+46.77	0.3619	+64.80
11	0.5802	0.3524	-39.26	0.4950	-14.68
12	0.3588	0.1466	-59.14	0.2319	-35.37
13	0.1745	0.0908	-47.97	0.0868	-50.26
14	0.6055	0.5604	-7.45	0.4963	-18.03
15	0.8877	0.9451	+6.47	0.8554	-3.64
16	0.3856	0.3094	-19.76	0.4823	+25.08
17	0.2360	0.1363	-42.25	0.1479	-37.33
18	0.7882	0.6419	-18.56	0.6662	-15.48
19	0.5141	0.4027	-21.67	0.4177	-18.75
20	0.1871	0.3997	+113.63	0.3016	+61.20
21	0.0152	0.0346	+127.63	0.0837	+450.66
22	0.0920	0.3644	+296.09	0.1399	+52.07
23	0.2328	0.4043	+73.67	0.4277	+83.72
24	0.3250	0.3177	-2.25	0.3951	+21.57
25	0.5943	0.2812	-52.68	0.3239	-45.50
26	0.2360	0.2312	-2.03	0.1034	-56.19
27	0.4634	0.3062	-33.92	0.3322	-28.31
28	0.0307	0.0306	-0.33	0.0142	-53.75
29	0.0314	0.2575	+720.06	0.3349	+966.56
30	0.2162	0.2164	+0.09	0.3832	+77.24
31	0.0500	0.0560	+12.00	0.0635	+27.00
32	0.4544	0.5968	+31.34	0.5803	+27.71
33	0.0220	0.0232	+5.45	0.0290	+31.82
34	0.2169	0.1989	-8.30	0.2299	+ 5.99
35	0.2267	0.3421	+50.90	0.4012	+76.97
36	0.0129	0.0286	+121.71	0.0406	+214.73
37	0.2563	0.2605	+1.64	0.2289	-10.69
38	0.2534	0.2300	-9.23	0.2079	-17.96
39	0.0006	0.0200	+3233.33	0.0085	+1316.67
40	0.2004	0.3230	+61.18	0.2708	+35.13
41	0.0015	0.4938	+32820.00	0.5261	+34973.33
42	0.2883	0.1346	-53.31	0.4216	+46.24
43	0.2996	0.1280	-57.28	0.1684	-43.79
44	0.0218	0.1019	+367.43	0.0952	+336.70
45	0.1506	0.1879	+24.77	0.2783	+84.79
46	0.3485	0.6087	+74.66	0.4719	+35.41
47	0.0967	0.0303	-68.67	0.3293	+240.54
48	0.3886	0.3418	-12.04	0.2954	-23.98
49	0.2066	0.1351	-34.61	0.1826	-11.62
50	0.3861	0.4312	+11.68	0.3978	+3.03
Average	0.2723	0.2815	+3.3	0.3033	+11.3

is a feedback aproach without requiring relevance information. Instead, an initial retrieval is performed, and the top-n ranked documents are all assumed to be relevant for obtaining expansion terms ($\overrightarrow{\mathrm{q}}_{feedback}$) as belows :

$$\overrightarrow{\mathrm{q}}_{feedback} = \frac{1}{|D_r|} \sum_{d_i \in D_r} \overrightarrow{d_i}$$

In this case, D_r is a set of documents ranked on the top in the initial retrieval and $\overrightarrow{d_i}$ is the vector representation of document d_i.

In the framework of the inference network (Xu and Croft 1996), the information need of the user is represented by multiple queries. Multiple queries means that an information need is represented by some different query representation. Experiments show that multiple query representations can produce better results than using one representation alone. However, how

to obtain these queries is not discussed in this model. Hence we try to find multiple query representations for the information structure derived from feedback information. In this way, the following three representations can be obtained :

- representation derived directly from the original query : $\overrightarrow{\mathrm{q}}_{original}$,
- representation obtained by our method : $\overrightarrow{\mathrm{q}}_{thesauri}$,
- representation derived from the retrieved documents of the previous run : $\overrightarrow{\mathrm{q}}_{feedback}$.

A linear combination of the three query representations is used to retrieve documents. However, we do not introduce additional parameters which are quite difficult to determine. Also we believe that the parameter values determined for some queries may not be suitable for some other queries because they are query dependent. Hence the simple combination we use is :

$$\overrightarrow{\mathrm{q}}_{original} + \overrightarrow{\mathrm{q}}_{thesauri} + \overrightarrow{\mathrm{q}}_{feedback}.$$

When using the relevance-feedback method, we used the top 30 ranked documents of the previous run of the original query to obtain $\overrightarrow{\mathrm{q}}_{feedback}$.

In order to evaluate the retrieval effectiveness of the new method, we carried out some experiments using TREC-7 collection to compare the retrieval effectiveness of the following methods using different combination of the query representations. Figure 5 shows 11-point interpolated precision using our method alone, pseudo-feedback alone, and the combination of our method and pseudo-feedback. Our method alone has better performance than the pseudo-feedback method, and the combination of our method and pseudo-feedback slightly better than our method alone.

Recently, Xu and Croft (1996) suggested a method called local context analysis, which also utilize the co-occurrence-based thesaurus and relevance feedback method. Instead of gathering co-occurrence data from the whole corpus, he gather it from the top-n ranked document. We carry out experiments in that we build the combined-thesauri based on the top-n ranked document, rather than the whole corpus. As can be seen in Figure 6, query expansion using the combined thesauri built from the top-n ranked document have a lower performance than query expansion using the combined thesauri built from the whole corpus.

7 Conclusions and Future Work

We have proposed the use of multiple types of thesauri for query expansion in information retrieval, give some failure analysis, and combining our method with pseudo-relevance feedback method. The basic idea underlying our method is that each type of thesaurus has

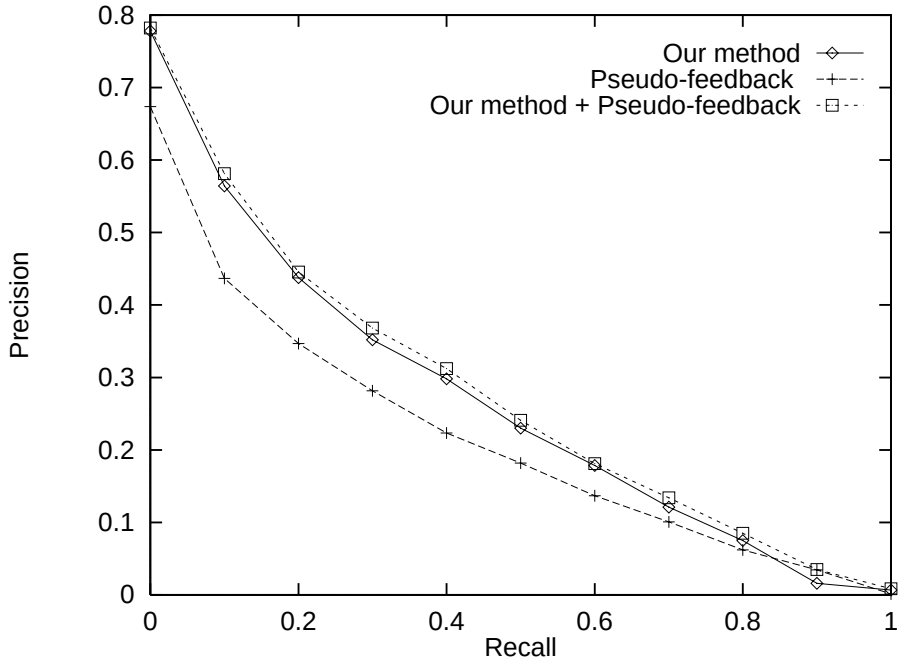

Fig. 5 The results of combining our method and pseudo-feedback

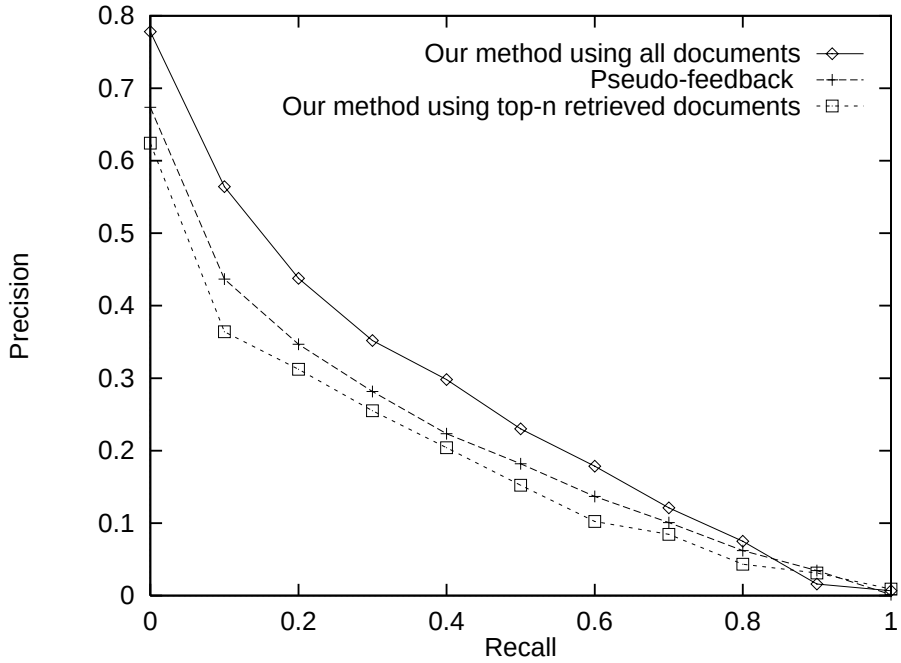

Fig. 6 The results of combined thesauri built from the top-n ranked document

different characteristics and combining them provides a valuable resource to expand the query. Misleading expansion terms can be avoided by designing a weighting term method in which the weight of expansion terms not only depends on all query terms, but also depends on their similarity values in all type of thesaurus.

Future research will include the use of parser with better performance, designing a general algorithm for automatically handling the negation statements, and also designing an effective algorithm for handling the multiple aspect contain in the query.

8 Acknowledgments

The authors would like to thank the anonymous referees for useful comments on the earlier version of this paper. We also thank Chris Buckley (SabIR Research) for support with SMART, Satoshi Sekine (New York University) for the Apple Pie Parser, Akitoshi Okumura (NEC C & C Media Lab.) for providing the computer environments in very preliminary experiments. This research is partially supported by JSPS project number JSPS-RFTF96P00502.

Reference

Aitchison, J. and Gilchrist, A. (1987). *Thesaurus Construction A Practical Manual*. Aslib.

Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addisson Wesley.

Blair, D. and Maron, M. (1985). "An evaluation of retrieval effectiveness." *Communications of the ACM*, *28*, 289–299.

Buckley, C. and Salton, G. (1994). "The Effect of Adding Relevance Information in a Relevance Feedback Environment." In *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval Conference*, pp. 292–300.

Buckley, C. and Salton, G. (1995). "Automatic Query Expansion using SMART: TREC-3." In *Proceedings of The Third Text Retrieval Conference*, pp. 69–80.

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

Chen, H., Schatz, B., Yim, T., and Fye, D. (1995). "Automatic Thesaurus Generation for an Electronic Community System." *Journal of American Society for Information Science*, *46*(3), 175–193.

Church, K. and Hanks, P. (1989). "Word Association Norms,Mutual Information and Lexicography." In *Proceedings of the 27nd Annual Meeting of the Association for Computational Linguistics*, pp. 76–83.

Crouch, C. J. (1990). "An Aproach to The Automatic Construction of Global Thesauri." *Information Processing and Management*, *26*(5), 629–640.

Crouch, C. and Yang, B. (1992). "Experiments in Automatic Statistical Thesaurus Construction." In *Proceedings of the Fifteenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval Conference*, pp. 77–82.

Ekmekcioglu, F. (1992). "Effectiveness of Query Expansion in Ranked-Output Document Retrieval Systems." *Journal of Information Science*, *18*, 139–147.

Fellbaum, C. (1998). *WordNet,An Electronic Lexical Database*. MIT Press.

Fox, E. A. (1980). "Lexical Relations Enhancing Effectiveness of Information Retrieval Systems." *SIGIR Forum*, *15*(3), 6–36.

Fox, E. A. (1990). *Virginia Disk One*. Blacksburg: Virginia Polytechnic Institute and State University.

Grafenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publisher.

Grefenstette, G. (1992). "Use of Syntactic Context to Produce Term Association Lists for Text Retrieval." In *Proceedings of the 15th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval Conference*, pp. 89–97.

Grossman, D. and Frieder, O. (1998). *Information Retrieval: Algorithms and Heuristics*. Kluwer Academic Publishers.

Hearst, M. A. (1994). "Multi-Paragraph Segmentation of Expository Text." In *Proceedings of 32th Annual Meeting of the Association for Computational Linguistics*, pp. 9–16.

Hearst, M. A. (1996). "Improving Full-Text Precision on Short Queries using Simple Constraints." In *Proceedings of the 5th Annual Symposium on Document Analysis and Information Retrieval (SDAIR)*.

Hearst, M. A. (1997). "TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages." *Computational Linguistics*, *23*(1), 33–64.

Hindle, D. (1990). "Noun Classification from Predicate-Argument Structures." In *Proceedings of 28th Annual Meeting of the Association for Computational Linguistics*, pp. 268–275.

Jing, Y. and Croft, B. (1994). "An Association Thesaurus for Information Retrieval." In *Proceedings of RIAO*, pp. 146–160.

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.

Leacock, C. and Chodorow, M. (1988). "Combining Local Context and WordNet Similarity for Word Sense Identification." In Fellbaum, C. (Ed.), *WordNet,An Electronic Lexical*

Database, pp. 265–283. MIT Press.

Lin, D. (1998). "Automatic Retrieval and Clustering of Similar Words." In *Proceedings of the COLING-ACL'98*, pp. 768–773.

Miller, G. (1990). "WordNet: An On-line Lexical Database." *Special Issue of the International Journal of Lexicography*, *3*(4).

Mitra, M., Singhal, A., and Buckley, C. (1998). "Improving Automatic Query Expansion." In *Proceedings of the 21th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR'98)*, pp. 206–214.

Paice, C. D. (1991). "A Thesaural Model of Information Retrieval." *Information Processing and Management*, *27*(5), 433–447.

Qiu and Frei, H. (1993). "Concept Based Query Expansion." In *Proceedings of the 16th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval Conference*, pp. 160–169.

Resnik, P. (1995). "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.

Richardson, R. and Smeaton, A. (1994). "Using WordNet for Conceptual Distance Measurement." In *Proceedings of the BCS-IRSG Colloqium*.

Richardson, R. and Smeaton, A. F. (1995). "Using WordNet in a Knowledge-Based Approach to Information Retrieval." Tech. rep. CA-0395, School of Computer Applications,Dublin City University.

Rila Mandala, Tokunaga, T., and Tanaka, H. (1999a). "Combining General Hand-Made and Automatically Constructed Thesauri for Information Retrieval." In *Proceedings of the 16th International Joint Conference on Artificial Intelligence (IJCAI'99)*, pp. 920–925.

Rila Mandala, Tokunaga, T., and Tanaka, H. (1999b). "Combining Multiple Evidence from Different Types of Thesaurus for Query Expansion." In *Proceedings of the 22th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99)*, pp. 191–197.

Rila Mandala, Tokunaga, T., and Tanaka, H. (1999c). "Complementing WordNet with Roget and Corpus-based Thesauri for Information Retrieval." In *Proceedings of the 9th European Chapter of the Association for Computational Linguistics (EACL'99)*, pp. 94–101.

Rila Mandala, Tokunaga, T., Tanaka, H., Okumura, A., and Satoh, K. (1999d). "Adhoc Retrieval Experiments Using WordNet and Automatically Constructed Thesauri." In *Proceedings of the Seventh Text REtrieval Conference (TREC-7)*, pp. 475–480. NIST

Special Publication.

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*. Prentice-Hall.

Salton, G. and Buckley, C. (1990). "Improving Retrieval Performance by Relevance Feedback." *Journal of American Society for Information Science*, *41*(4), 288–297.

Salton, G. and McGill, M. (1983). *An Introduction to Modern Information Retrieval*. McGraw-Hill.

Schutze, H. and Pederson, J. (1994). "A Cooccurrence-Based Thesaurus and Two Applications to Information Retrieval." In *Proceedings of RIAO Conference*, pp. 266–274.

Schutze, H. and Pederson, J. (1997). "A Cooccurrence-Based Thesaurus and Two Applications to Information Retrieval." *Information Processing and Management*, *33*(3), 307–318.

Sekine, S. and Grishman, R. (1995). "A Corpus-based Probabilistic Grammar with Only Two Non-terminals." In *Proceedings of the International Workshop on Parsing Technologies*.

Smeaton, A. and Berrut, C. (1996). "Thresholding Postings Lists,Query Expansion by Word-Word Distances and POS Tagging of Spanish Text." In *Proceedings of The Fourth Text Retrieval Conference*.

Stairmand, M. (1997). "Textual Context Analysis for Information Retrieval." In *Proceedings of the 20th ACM-SIGIR Conference*, pp. 140–147.

Voorhees, E. M. (1988). "Using WordNet for Text Retrieval." In Fellbaum, C. (Ed.), *WordNet,An Electronic Lexical Database*, pp. 285–303. MIT Press.

Voorhees, E. (1994). "Query Expansion using Lexical-Semantic Relations." In *Proceedings of the 17th ACM-SIGIR Conference*, pp. 61–69.

Voorhees, E. and Harman, D. (1999). "Overview of the Seventh Text retrieval Conference (TREC-7)." In *Proceedings of the Seventh Text REtrieval Conference*. NIST Special Publication.

Xu, J. and Croft, B. (1996). "Query Expansion Using Local and Global Document Analysis." In *Proceedings of the 19th ACM-SIGIR Conference*, pp. 4–11.

Rila Mandala: He is a lecturer in Department of Informatics, Bandung Institute of Technology, Indonesia since 1992. He received the B.S. degree in informatics from Bandung Institute of Technology, Indonesia and M.Eng. degree in computer science from Tokyo Institute of Technology, Japan, in

1992 and 1996, respectively. Currently, he is a doctoral student of Department of Computer Science, Tokyo Institute of Technology. His current research interests are information retrieval, computational linguistics, and natural language processing.

Takenobu Tokunaga: He is an associate professor of Graduate School of Information Science and Engineering, Tokyo Institute of Technology. He received the B.S. degree in 1983 from Tokyo Institute of Technology, the M.S and the Dr.Eng. degrees from Tokyo Institute of Technology in 1985 and 1991, respectively. His current interests are computational linguistics and information retrieval.

Hozumi Tanaka: He is a professor of Department of Computer Science, Tokyo Institute of Technology. He received the B.S. degree in 1964 and the M.S. degree in 1966 from Tokyo Institute of Technology. In 1966 he joined in the Electro Technical Laboratories, Tsukuba. He received the Dr.Eng. degree in 1980. He joined in Tokyo Institute of Technology in 1983. He has been engaged in artificial intelligence and natural language processing research.

(Received October 25, 1999)
(Revised December 6, 1999)
(Accepted January 7, 2000)