Corpus-dependent Association Thesauri for Information Retrieval

Hiroyuki Kaji[*1], Yasutsugu Morimoto[*1], Toshiko Aizono[*1], and Noriyuki Yamasaki[*2]

[*1] Central Research Laboratory, Hitachi, Ltd.
1-280 Higashi-Koigakubo, Kokubunji-shi
Tokyo 185-8601, Japan

[*2] Software Division, Hitachi, Ltd.
549-6 Shinano-cho, Totsuka-ku, Yokohama-shi
Kanagawa 244-0801, Japan

{kaji, morimoto, aizono}@crl.hitachi.co.jp, yamasa_n@soft.hitachi.co.jp

Abstract

This paper presents a method for automatically generating an association thesaurus from a text corpus, and demonstrates its application to information retrieval. The thesaurus generation method consists of extracting terms and co-occurrence data from a corpus and analyzing the correlation between terms statistically. A new method for disambiguating the structure of compound nouns, which is a key component for term extraction, is also proposed. The automatically generated thesaurus is effectively used as a tool for exploring information. A thesaurus navigator having novel functions such as term clustering, thesaurus overview, and zooming-in is proposed.

1 Introduction

A thesaurus plays essential roles in information retrieval systems. In particular, a domain-specific thesaurus greatly improves the effectiveness of information retrieval. However, we are confronted with the difficult problem of how to construct and maintain a domain-specific thesaurus. The goal of our present research is to establish a method for automatically generating a thesaurus from a text corpus of a domain and demonstrate its application to information retrieval.

Thesauri are classified into taxonomy-type thesauri and association thesauri. There has been various research on the extraction of taxonomic information from a corpus, including extraction of hyponyms by using linguistic patterns (Hearst 1992) and extraction of synonyms based on the similarity of sets of co-occurring words (Ruge 1991; Grefenstette 1992). However, the performance of these methods is limited, and they should be considered as aids to augment handmade thesauri. In contrast, an association thesaurus, that is a collection of pairs of semantically associated terms, can be possibly generated from a corpus entirely automatically. Word association norms based on co-occurrence information have been proposed by (Church and Hanks 1990). Here we focus on the automatic generation of an association thesaurus.

Association thesauri are as useful as taxonomy-type thesauri in information retrieval. The improvement of retrieval effectiveness by using an association thesaurus has been reported by a number of papers (Jing and Croft 1994; Schutze and Pedersen 1994). We propose to use a corpus-dependent association thesaurus interactively.

2 Automatic Generation of an Association Thesaurus from a Corpus

2.1 Outline of the thesaurus generation method

The proposed thesaurus generation method consists of term extraction, co-occurrence data extraction, and correlation analysis, as shown in Fig. 1.

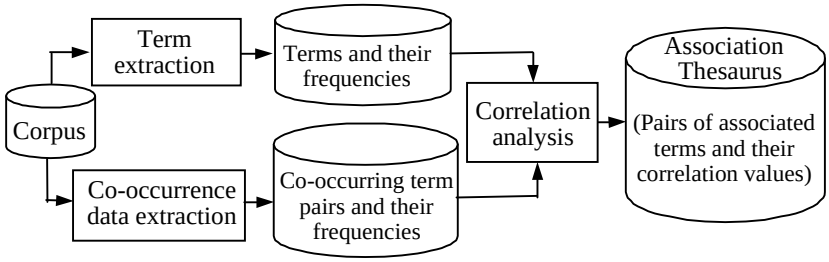


Fig. 1. Automatic generation of a thesaurus from a corpus.

2.1.1 Term extraction

A thesaurus should consist of terms, each representing a domain-specific concept. Most of the terms representing important concepts are nouns, simple or compound, that frequently occur in the corpus. Therefore, we extract both simple nouns and compound nouns whose occurrence frequencies exceed a predetermined threshold. We also use a list of stop words since frequently occurring nouns are not always terms.

Compound nouns are identified by a pattern matching method using a part-of-speech sequence pattern. Naturally, the pattern is language specific. The following is a pattern for Japanese compound nouns:

COMP_NOUN := { { PREFIX } NOUN + SUFFIX } { PREFIX } NOUN +

A problem in extracting compound nouns is that a word sequence matched to the above pattern, which actually defines just the type of noun phrase, is not always a term. We filter out some kind of non-term noun phrases by using a list of stop words for the first and last elements of compound nouns. Stop words for the first element of compound nouns include referential nouns (e.g. *jouki* (above-mentioned)) and determiner nouns (e.g. *kaku* (each)). Stop words for the last element of compound nouns include time/place nouns (e.g. *nai* (inside)) and relational nouns (e.g. *koyuu* (peculiar)).

Another important problem we are confronted with in term extraction is the structural ambiguity of compound nouns. For our purpose, we need to extract non-maximal compound nouns as well as maximal compound nouns. Here a non-maximal compound noun means one that occurs as a part of a larger compound noun, and a maximal compound noun means one that occurs not as a part of a larger compound noun. We must disambiguate the structure of compound nouns to correctly extract non-maximal compound nouns. We have developed a statistical disambiguation method, the detail and evaluation of which are described in 2.2.

2.1.2 Co-occurrence data extraction

Our purpose is to collect pairs of semantically or contextually associated terms, no matter what kind of association. So we extract co-occurrence in a window. That is, every pair of terms occurring together within a window is extracted as the window is moved through a text. The window size can be specified rather arbitrarily. Considering our purpose, the window should accommodate a few sentences. At the same time, the window size should not be too large from the viewpoint of computational load. Therefore, 20 to 50 words, excluding function words, seems to be an appropriate value.

Note that we filter out any pair of words co-occurring within a compound noun. If such pairs were included in co-occurrence data, they would show high correlation. However, they would be redundant because compound nouns are treated as entities in our thesaurus.

2.1.3 Correlation analysis

As a correlation measure between terms, we use mutual information (Church and Hanks 1990). The mutual information between terms t_i and t_j is defined by the following formula:

$$MI(t_i, t_j) = \log_2 \frac{g(t_i, t_j) \Big/ \sum_{i,j} g(t_i, t_j)}{\left\{ f(t_i) \Big/ \sum_i f(t_i) \right\} \cdot \left\{ f(t_j) \Big/ \sum_j f(t_j) \right\}},$$

where $f(t_i)$ is the occurrence frequency of term t_i, and $g(t_i,t_j)$ is the co-occurrence frequency of terms t_i and t_j. A maximum number of associated terms for each term is predetermined as well as a threshold for the mutual information, and associated terms are selected based on the descending order of mutual information.

Mutual information involves a problem in that it is overestimated for low-frequency terms (Dunning 1993). Therefore, we determine whether two terms are related to each other by a log-likelihood ratio test, and we filter out pairs of terms that do not pass the test.

2.2 Disambiguation of compound noun structure

2.2.1 Disambiguation based on corpus statistics

Our disambiguation method is described below for the case of a compound noun consisting of three elements. A compound noun $W_1W_2W_3$ has two possible structures: $W_1(W_2W_3)$ and $(W_1W_2)W_3$. We determine its structure based on the occurrence frequencies of maximal compound nouns as follows: If the maximal compound noun W_2W_3 occurs more frequently than the maximal compound noun W_1W_2, then the

Table 1 Global-statistics-based disambiguation vs. local-statistics-based disambiguation.
(a) Examples

Maximal compound noun	Freq.	Global-statistics-based		Local-statistics-based	
		Structure	Freq.	Structure	Freq.
Deta shori shisutemu (data processing system)	478	(*Deta shori*) *shisutemu*	478	(*Deta shori*) *shisutemu*	368
				Deta (*shori shisutemu*)	110
Deta tsushin seigyo souchi (Data communication controller)	94	*Deta* (*tsushin seigyo souchi*)	94	(*Deta tsushin*) *seigyo souchi*	40
				Deta (*tsushin seigyo souchi*)	54
Kaisen seigyo purosessa (Line control processor)	54	*Kaisen* (*seigyo purosessa*)	54	(*Kaisen seigyo*) *purosessa*	54

(b) Summary of disambiguation results

	Global-statistics-based	Local-statistics-based
Correct structure	12,565 words (62.0%)	14,921 words (73.7%)
Incorrect structure	7,688 words (38.0%)	5,332 words (26.3%)
Total	20,253 words (100%)	20,253 words (100%)

(Note: Numbers of words are occurrence-based.)

structure $W_1(W_2W_3)$ is preferred. On the contrary, if the maximal compound noun W_1W_2 occurs more frequently than the maximal compound noun W_2W_3, then the structure $(W_1W_2)\ W_3$ is preferred.

The generalized disambiguation rule is as follows: If a compound noun CN includes two compound noun candidates CN_1 and CN_2, which are incompatible with each other, and the maximal compound noun CN_1 occurs more frequently than the maximal compound noun CN_2, then a structure of CN including CN_1 is preferred to a structure of CN including CN_2.

We have two alternatives regarding the range where we count occurrence frequencies of maximal compound nouns. One is global-statistics which means that frequencies are counted in the whole corpus and they are used to disambiguate all compound nouns in the corpus. The other is local-statistics which means that frequencies are counted in each document in the corpus and they are used to disambiguate compound nouns in the corresponding document.

2.2.2 Evaluation: Global-statistics vs. local-statistics

We evaluated both the global-statistics-based disambiguation and the local-statistics-based disambiguation by using a 23.7-M Byte corpus consisting of 800 patent documents. Table 1(a) shows comparative examples of these methods. Evaluation results for the 200 highest-frequency maximal compound nouns consisting of three or more words are summarized in Table 1(b). They prove that the local-statistics-based disambiguation method is superior to the global-statistics-based disambiguation method. Note that in the local-statistics-based disambiguation method, we resorted to the global-statistics when local-statistics were not available. The percentage of cases the local-statistics were not available was 25.1 percent.

(Kobayasi et al. 1994) proposed a disambiguation method using collocation information and semantic categories, and reported that the structure of compound nouns was disambiguated at 83% accuracy. Note that their accuracy was calculated for compound nouns including unambiguous compound nouns, i.e. those consisting of only two words. If it were calculated for compound nouns consisting three or more words, it would be less than that of our method. Thus, we can conclude that our local-statistics-based method compares quite well with rather sophisticated previous methods.

2.3 Prototype and an experiment

We implemented a prototype thesaurus generator in which the local-statistics-based method was used to disambiguate the structure of compound nouns. Using this thesaurus generator, we got a thesaurus consisting of 38,995 terms from a 61-M Byte corpus consisting of almost 48,000 articles in the financial pages of a Japanese newspaper. In this experiment, the threshold for occurrence frequencies of terms in the term extraction step was set to 10, and the window size in the co-occurrence data extraction step was set to 25.

The above run took 5.4 hours on a HP9000 C200 workstation. The throughput is tolerable from a practical point of view. We should also note that a thesaurus can be updated as efficiently as it can be initially generated. Because we can run the first two steps (extraction of terms and extraction of co-occurrence data) in accumulative fashion, and we only need to run the third step over again when a considerable amount of terms and co-occurrence data are accumulated.

3 Navigation in an Association Thesaurus

3.1 Purpose and outline of the proposed thesaurus navigator

A big problem with today's information retrieval systems based on search techniques is that they require users, who may not know exactly what they are looking for, to explicitly describe their information needs. Another problem is that mismatched vocabularies between users and the corpus would bring poor retrieval results. To solve these problems, we propose a corpus-dependent association-thesaurus navigator enabling users to efficiently explore information through a corpus.

Users' requirements are summarized as follows:

- They want to grasp the overall information structure of a domain.
- They want to know what topics or sub-domains are contained in the corpus.
- They want to know terms that appropriately describe their vague information needs.

To meet the above requirements, our proposed thesaurus navigator has novel functions such as clustering of related terms, generation of a thesaurus overview, and zoom-in on a sub-domain of interest. A conceptual image of thesaurus navigation using these functions is shown in Fig. 2. A typical information-exploration session proceeds as follows.

At the beginning, the system displays an overview of a corpus-dependent thesaurus so that users can easily enter the information space of the corpus. The overview is a kind of summary of the corpus. It consists of clusters of generic terms of the domain, and makes it easy to understand what topics or sub-domains are contained in the corpus. Looking at the thesaurus overview, the users can select one or a few term clusters they have interest in, and the screen will zoom in on the cluster(s). The zoomed view consists of a number of clusters, each including more specific terms than those in the overview. Users can repeat this zoom-in operation until they reach term clusters representing sufficiently specific topics.

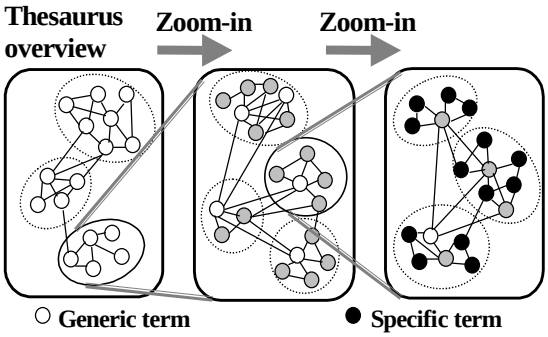


Fig. 2. Conceptual image of thesaurus navigation.

3.2 Functions of the thesaurus navigator

3.2.1 Clustering of related terms

We made a preliminary experiment to evaluate standard agglomerative clustering algorithms including the single-linkage method, the complete-linkage method, and the group-average-linkage method (El-Hamdouchi and Willett 1989). Among them, the group-average-linkage method resulted in the best results. However, several potential clusters tended to merge into a large one when we repeated the merge operation until a predetermined number of clusters were obtained. Accordingly, we use the group-average-linkage method with an upper limit on the size of a cluster.

3.2.2 Generation of a thesaurus overview

Our method for generating a thesaurus overview consists of major-term extraction and term clustering. The major-term extracting algorithm, which is carried out beforehand in batch mode, is described below. See 3.2.1 for the term clustering algorithm.

An overview of the thesaurus should consist of generic terms included in the corpus. However, we do not have a definite criterion for generic terms. So we collect major terms from the corpus as follows. The number of major terms,

denoted by M below, was set to 300 in the prototype.

i) Determine a characteristic term set for each document.

Calculate the weight $w_{i,j}$ of term t_j for document d_i according to the tf-idf (term frequency - inverse document frequency) formula. Then select the first $m(i)$ terms in the descending order of $w_{i,j}$ for each document d_i, where $m(i)$, the number of characteristic terms for document d_i, is set to 20% of the total number of distinct terms in d_i. It is also limited to between 5 and 50.

ii) Select major terms in the corpus.

Select the first M terms in the descending order of the frequency of being contained in the characteristic term sets.

3.2.3 Zoom-in on a term cluster of interest

Our method for zooming in on a term cluster consists of term-set expansion and term clustering. The term-set expanding algorithm is described below. See 3.2.1 for the term clustering algorithm.

A user-specified term set $T_o = \{t_1, t_2, \ldots, t_m\}$ is expanded into a term set T_e consisting of M terms as follows. M was set to 300 in the prototype.

i) Set the initial value of T_e to T_o.

ii) **While** $|T_e| < M$ **for** $i = 1, 2, \ldots$ **do**;

While $|T_e| < M$ **for** $j = 1, 2, \ldots, m$ **do**;

Add the term having the i-th highest correlation with t_j to T_e;

end;

end;

The reason why the above-described procedure implements the zoom-in is that generic terms tend to have higher correlation with semi-generic terms than with specific terms. Assuming that high-frequency terms are generic and low-frequency terms are specific, we examined the distribution of terms by the distance from the major terms and the average occurrence frequency of terms for each distance. Here the distance is the length of the shortest path in a graph that is obtained by connecting every pair of associated terms with an edge. Table 2 shows the results for the example thesaurus mentioned in 2.3. According to it, the average occurrence frequency decreases with the distance from the major terms. Therefore, starting from an overview, our method is likely to produce more and more specific views.

3.3 Prototype and an experiment

We developed a prototype as a client/server system. The thesaurus navigator is available on WWW browsers. It also has an interface to text-retrieval engines, through which a term cluster is transferred as a query.

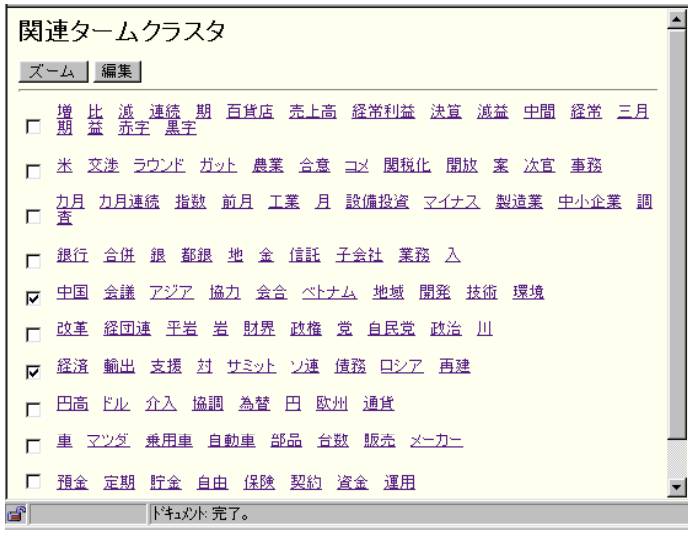
Test use was made with the example thesaurus mentioned in 2.3. The response time for the zoom-in operation during the navigation sessions was about 8 seconds. This is acceptable given the rich information provided by the clustered view. Note that the response time is almost independent of the size of the thesaurus or corpus, because the number of terms to be clustered is always constant, as described in 3.2.2 and 3.2.3.

An example from navigation sessions is shown in Fig. 3. It demonstrates the usefulness of the corpus-dependent thesaurus navigation as a front-end for text retrieval. The effectiveness of our thesaurus navigator is summarized as follows.

- Improved accessibility to text retrieval systems:
  Users are not necessarily required to input terms to describe their information need. They need only select from among terms presented on the screen. This makes text retrieval systems accessible even for those having vague information needs, or those unfamiliar with the domain.
- Improved navigation efficiency:
  The unit of users' cognition is a topic rather than a term. That is, they can recognize a topic from a cluster of terms at a glance. Therefore, they can efficiently navigate through an information space.

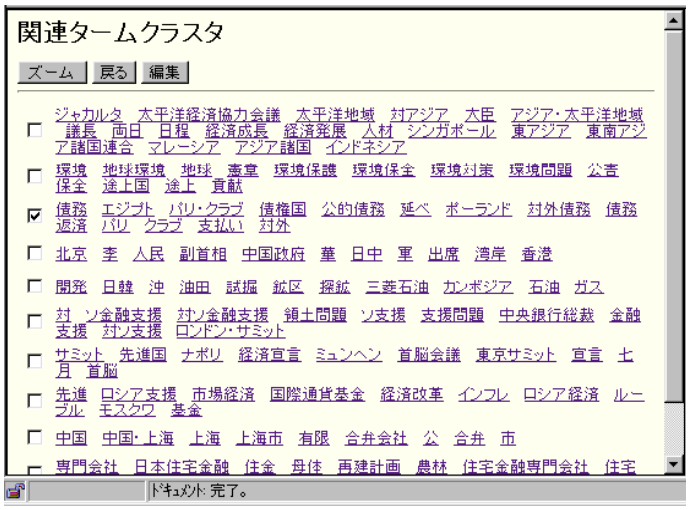
Table 2 Distribution of terms by distance from major terms.

Distance from major terms	0	1	2	3	4	5	∞
Number of terms	245	4278	23832	10408	149	1	82
Average occurrence frequency	2642.1	318.6	61.9	10.6	7.7	9.0	-



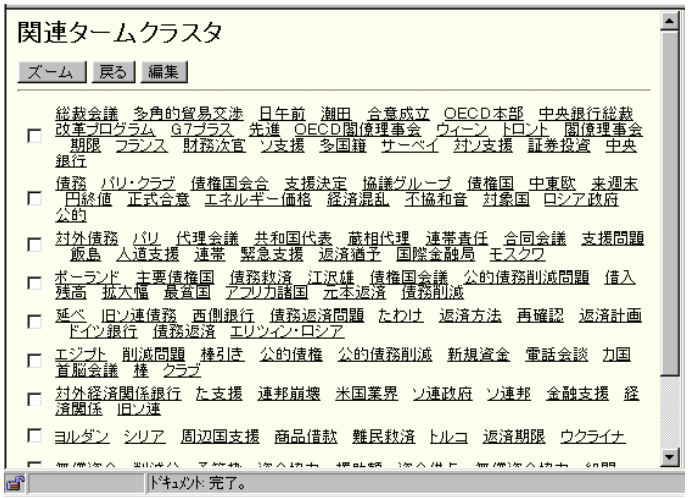
(a) Thesaurus overview

An overview of the thesaurus was displayed. Then the user selected the fifth and seventh clusters which he was interested in: {China, conference, Asia, cooperation, meeting, Vietnam, region, development, technology, environment}, and {economy, export, aid, toward, summit, Soviet Union, debt, Russia, reconstruction}. This means that the user was interested in "development assistance to developing countries or areas".



(b) Zoom-in

The fifth and seventh clusters from (a) were shown close up, and clusters indicating more specific domains were presented. The user could understand which topics the respective clusters suggested: "Economic assistance for the development of the Asia-Pacific region", "Global environmental problems", "International debt problems", "Matters on China", "Energy resource development", and so on. Since he was especially interested in "International debt problems", he selected the third cluster {debt, Egypt, Paris Club, creditor nation, official debt, deferred, Poland, foreign debt, reimbursement, Paris, club, payment, foreign}.



(c) Further zoom-in

The third cluster from (b) was shown close up. The resulting screen gave the user a choice of many specific terms relevant to "International debt problems", although not all of the clusters indicated specific topics. The user was able to retrieve documents by simply selecting terms of interest from those displayed on the screen.

Fig. 3. Example of thesaurus navigation.

4 Comparison with related work

Let us make a brief comparison with related work. Both scatter/gather document clustering (Cutting et al. 1992; Hearst and Pedersen 1996) and Kohonen's self-organizing map (Lin et al. 1991; Lagus et al. 1996; Kohonen 1998) enable exploration through a corpus. While they treat a corpus as a collection of documents, we treat it as a collection of terms. Therefore our method can elicit finer information structure than these previous methods, and moreover, it can be applied to a corpus that includes multi-topic documents. Our method compares quite well with the previous methods for throughput and response time.

5 Conclusion

We demonstrated the feasibility of automatic generation of an association thesaurus from a corpus. The proposed thesaurus generation method consists of extracting terms and co-occurrence data from a corpus and analyzing the correlation between terms statistically. As a component technology for thesaurus generation, a method for disambiguating the structure of compound nouns based on corpus statistics was developed and evaluated.

We also demonstrated the information retrieval application of an automatically generated association thesaurus. A thesaurus navigator having novel functions such as term clustering, thesaurus overview, and zooming-in was developed. An experiment with an association thesaurus generated from a newspaper article corpus proved that the thesaurus navigator allows us to efficiently explore information through a text corpus even when our information needs are vague.

Acknowledgements: This research was supported in part by the Next-generation Digital Library System R&D Project of MITI (Ministry of International Trade and Industry), IPA (Information-technology Promotion Agency), and JIPDEC (Japan Information Processing Development Center). We thank Mainichi Newspapers, Ltd. for permitting us to use the CD-ROMs of the '91, '92, '93, '94 and '95 *Mainichi Shimbun* for the experiment.

References

Church, K. W., and P. Hanks. 1990. Word association norms, mutual information, and lexicography. Computational Linguistics, 16(1): 22-29.

Cutting, D. R., D. R. Karger, J. O. Pedersen, and J. W. Tukey. 1992. Scatter/gather: A cluster-based approach to browsing large document collections. Proc. ACM SIGIR '92, pp. 318-329.

Dunning, T. 1993. Accurate methods for the statistics of surprise and coincidence. Computational Linguistics, 19(1): 61-74.

El-Hamdouchi, A., and P. Willett. 1989. Comparison of hierarchical agglomerative clustering methods for document retrieval. The Computer Journal, 32(3): 220-227.

Grefenstette, G. 1992. Use of syntactic context to produce term association lists for text retrieval. Proc. ACM SIGIR '92, pp. 89-97.

Hearst, M. A. 1992. Automatic acquisition of hyponyms from large text corpora. Proc. COLING '92, pp. 539-545.

Hearst, M. A., and J. O. Pedersen. 1996. Reexamining the cluster hypothesis: Scatter/gather on retrieval results. Proc. ACM SIGIR '96, pp. 76-84.

Jing, Y., and W. B. Croft. 1994. An association thesaurus for information retrieval. Proc. RIAO '94, Conf. on Intelligent Text and Image Handling, pp. 146-160.

Kobayasi, Y., T. Tokunaga, and H. Tanaka. 1994. Analysis of Japanese compound nouns using collocational information. Proc. COLING '94, pp. 865-869.

Kohonen, T. 1998. Self-organization of very large document collections: State of the art. Proc. 8th Int'l Conf. on Artificial Neural Networks, vol. 1, pp. 65-74.

Lagus, K., T. Honkela, S. Kaski, and T. Kohonen. 1996. Self-organizing maps of document collections: a new approach to interactive exploration. Proc. 2nd Int'l Conf. on Knowledge Discovery and Data Mining, pp. 238-243.

Lin, X., D. Soergel, and G. Marchionini. 1991. A self-organizing semantic map for information retrieval. Proc. ACM SIGIR '91, pp. 262-269.

Ruge, G. 1991. Experiments on linguistically based term associations. Proc. RIAO '91, Conf. on Intelligent Text and Image Handling, pp. 528-545.

Schutze, H., and J. O. Pedersen. 1994. A cooccurrence-based thesaurus and two applications to information retrieval. Proc. RIAO '94, Conf. on Intelligent Text and Image Handling, pp. 266-274.