Pacific Association for Computational Linguistics

VECTOR SPACE MODEL BASED ON SEMANTIC ATTRIBUTES OF WORDS

Satoru Ikehara *, Jin'ichi Murakami * , Yasuhiro Kimoto * , Teturou Araki **

* *ikehara, murakami, kimoto@ike.tottori-u.ac.jp*
Faculty of Engineering, Tottori University Minami 4-101, Tottori-city, 680-8552 Japan
* *araki araki@knowipc.fuee.fukui-u.ac.jp*
Department of Human and Artificial Intelligent Systems Fukui University Bunkyou 3-9-1, Fukui, Fukui 910-8507, Japan

In order to reduce the dimension of VSM (Vector Space Model) for information retrieval and clustering, this paper proposes a new method, Semantic-VSM, which uses the Semantic Attribute System defined by "*A-Japanese-Lexicon*" instead of literal words used in conventional VSM.

The attribute system consists of a tree structure with 2,710 attributes, which includes 400 thousand literal words. Using this attribute system, the generalization of vector elements can be performed easily based on upper-lower relationships of semantic attributes, so that the dimension can easily be reduced at very low cost. Synonyms are automatically assessed through semantic attributes to improve the recall performance of retrieval systems.

Experimental results applying it to BMIR-J2 database of 5,079 newspaper articles showed that the dimension can be reduced from 2,710 to 300 or 600 with only a small degradation in performance. High recall performance was also shown compared with conventional VSM.

1. INTRODUCTION

With the increasing availability of information in electronic form, it becomes more important and feasible to have automatic methods to retrieve such information. In addition to the conventional method by Key Words, many new methods, such as full-text search, passage retrieval, contents retrieval and VSM (Vector Space Model) have been investigated.

Among these methods, VSM is one promising method for improving the performance of information retrieval as well as clustering. However, conventional VSM uses so many words per vector element that similarity calculation requires much time. When the query sentence includes only a few words pertaining to the vector elements, the query vector becomes too sparse to find the relevant documents.

In order to resolve these problems, many researches have been conducted. The most simple way to reduce the vector dimension is the selection of elements based on the value of $tf \cdot idf$(Salton, McGill 1983). Hierarchical classification analyses are frequently used for term and document clustering (Jardin et.al, 1971).

In the case of VMS, it has been assumed that the meaning of the words which represent the bases of vector are independent from each other, however, this assumption does not hold in actual documents. Then, in order to reduce the number of dimension, KL method (Borco and Bernick 1963) and LSI method (Deerwester et al. 1990, Faloutsos and Lin 1995, Golub and Loan 1996)were proposed where new bases were generated by linear combination of vector bases.

Semantic similarities between bases were considered in KL method and the vectors which represent each cluster were selected as the new vector bases. On the other hand, LSI(Latent Semantic Indexing) tries to find the new meanings behind plural words used for bases. It finds the new bases from the matrix composed of specific vectors by using SVD(Singular Value Decomposition, Golub and Kahan 1965) method. This method was applied slso to a numerical database(Jiang et al. 1999).

LSI is an attractive method that can reduce the dimension without decreasing the performance of information retrieval. However, the calculation of SVD requires much time to

© 2001 Pacific Association for Computational Linguistics

apply it to a large number of documents. In some cases, vector bases are determined from the limmitted number of the documents(Deerwester et al. 1990).

In addition to the above, a pseudo-feed back method (named as Two Stage ad-hoc retrieval) was also proposed(Burkley et al. 1996, Kwock and Chan 1998).

By the way, Mining Term Association was known as the learning method to acquire the semantic relations between words and applied to the documents from the Internet(Lin et al. 1998). However, it is difficult automatically to determine the semantic relation of words at high accuracy.

In order to resolve these problems, this paper proposes a new method using semantic attributes as vector elements instead of literal terms, which can easily reduce the dimension without decreasing the performance.

In this method, the **Semantic Attribute System** defined by "*A-Japanese-Lexicon*" (Ikehara et. al, 1997) is used. The semantic usage of Japanese words was hierarchically classified from the view point of "is-a" and "has-a" relationship into 2,710 categories called "Semantic Attribute". The semantic usage of 400 thousand Japanese words was defined using this system. Therefore, the meanings of most of the Japanese words used in the documents can be represented by Semantic Attributes; the similarity of meanings between a query sentence and the documents in the database are assessed through these Semantic Attributes to improve the "recall" performance. It is expected that the vector dimension can easily be reduced using upper-lower relations between Semantic Attributes.

In this paper, experiments in information retrieval are conducted applying our method to TREC test collection "BMIR-J2" (Kitani 1998) to evaluate the efficiency compared to conventional VSM.

2. SEMANTIC-VECTOR SPACE MODEL

2.1. Conventional Model

(1) Meanings of Sentence and Document

Conventional VSM assumes that meanings of a sentence and a document are represented by a set of words used in them. The set of words is represented by specific vector V as follows:

$$V = (w_1, w_2, \cdots w_i, \cdots w_n) \tag{1}$$

Here, i $(1 \leq i \leq n)$ is the number of words which are used to represent the meanings of sentences.

As for words to be used for vector elements, similarly to the information retrieval system which uses controlled Key Words, important words are statistically selected by some conventional method, such as "$tf \cdot idf$" from all of the documents in the database. The values of weight w_i are usually determined dependent on the frequency of the appearance of word $\#i$.

Here, we call the specific vector given by

(1) as "**Word-Vector**" and the VSM which uses this type of the specific vector as "**W-VSM**" (**Word-Vector Space Model**).

(2) Semantic Similarity of Documents

The semantic similarity $sim(D_i, D_j)$ between two documents, D_i and D_j, is defined by the inner product of the specific vector V_i and V_j.

$$sim(D_i, D_j) = V_i \cdot V_j \tag{2}$$

This relation is used for clustering documents and information retrieval.

2.2. Semantic-VSM

(1) Semantic Vector

Here, we propose a new method in which, instead of literal words, meanings of words are used as the vector elements. In this method, the meanings of all of the Japanese words are classified into k categories and they are used as the bases of the specific vector.

Here, let S_i represent the weight of all the words in document D_j which have the meaning of $\#i$, the specific vector V_j for the document D_j, is written as;

$$V_j = (S_1, S_2, \cdots S_i, \cdots S_k) \tag{3}$$

Similarly to Word-VSM, there are many ways to determine the values of weight Wi. We use the values of $tf \cdot idf$.

Hereafter, we call the specific vector given by (3) as "**Semantic-Vector**" and the VSM which uses this type of specific vector as "**S-VSM**" (**Semantic Vector Space Model**).

(2) Semantic Attribute System for Japanese

In order to implement Semantic-VSM, we use the Semantic Attribute System which was recently proposed in "A-Japanese Lexicon"(Ikehara et. al, 1997). This system is shown in part in Fig.**??**.

In this system, semantic usage of Japanese words is classified into 2,710 attributes and the relationships among them, namely, "is-a" relation and "has-a" relation, are represented by a tree with 12 levels. A semantic word dictionary is also given in this lexicon, where the semantic usage of 400 thousand Japanese words is defined by using their Semantic Attributes.

Thus, if the frequency of words used in documents is obtained, the values of vector elements S_i in eq.(3) can easily be calculated by summing up such words that have the meaning $\#i$.

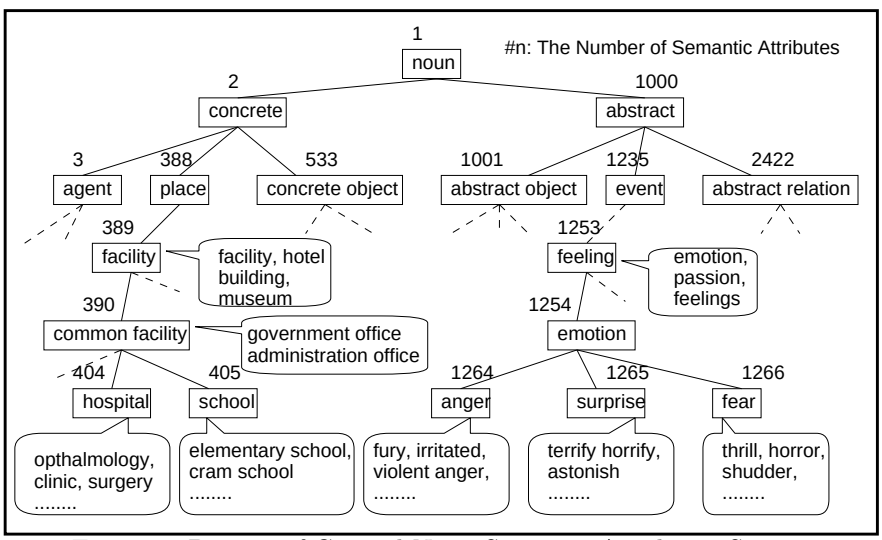

Figure 1: Portion of General Noun Semantic Attributes System

2.3. Characteristics of S-VSM

Let us consider the characteristics of S-VSM compared to conventional W-VSM.

(1) Possibility of Reducing Dimension

VSM assumes that the bases of specific vectors are independent from each other. However, this assumption does not hold in actual documents. There are many axes which are mutually dependent or not used. Therefore, it is very important to find the latent semantic structure.

LSI is known as a useful method to find such structures and to reduce the vector dimension. However, *Singular Value Decomposition* poses a problem, in that it uses too much computing time to apply it to a large number of documents.

Compared to this method, in our method, generalization of vector elements can be performed easily using the semantic relations between Semantic Attributes to reduce the dimension.

In the generalization (see below) of vector space, the attributes which are not so important are deleted, but their value is added to an upper node in the tree. Thus, the deleted attributes also contribute to a certain extent to the performance of information retrieval, and it is expected that the dimension can be reduced greatly without decreasing the performance of information retrieval.

(2) Possibility of Improving Recall Factor

In conventional W-VSM, a word never contributes to performance if it is not used as the vector element, even if it appears in the documents. Words that are literally different are treated as words with different meanings.

On the other hand, in S-VSM, 400 thousand words contribute to the performance of information retrieval through 2,710 Semantic Attributes. The meanings of most of the words used in documents, regardless of whether they are synonyms or not, reflect specific vectors and the recall performance is expected to improve compared to conventional methods.

3. IRREDUCIBLE MINIMUM VECTOR

In this section, we describe the method for finding the minimum set of bases for the semantic vector.

3.1. Generalization of Vector Elements

In order to reduce the dimension, we use "*generalization methods.*" Generalization is an inductive inference method to find rules from examples. In this paper, "*generalization*" means degenerating the semantic attributes which have little effect in information retrieval into the upper attribute. The weight of the attributes which are deleted from bases are added to the weight of the upper node.

Here, we consider the following two kinds of generalizations:

(1)Generalization by Meaning Granularity

As mentioned before, the semantic attribute system has a tree structure of 12 degrees. Lower semantic attributes have smaller granularity of meaning. Then, if we regard the degree of the semantic attributes as their granularity, the attributes lower than the threshold degree

are generalized. Fig.2 shows a case in which the attributes lower than or equal to 8 degrees are generalized.

(2) Generalization by Meaning Weight

The attributes with small weight will become the target of generalization because they have small effects on information retrieval.

Fig.2 shows a case in which the semantic attributes with a weight smaller than 5 are generalized.

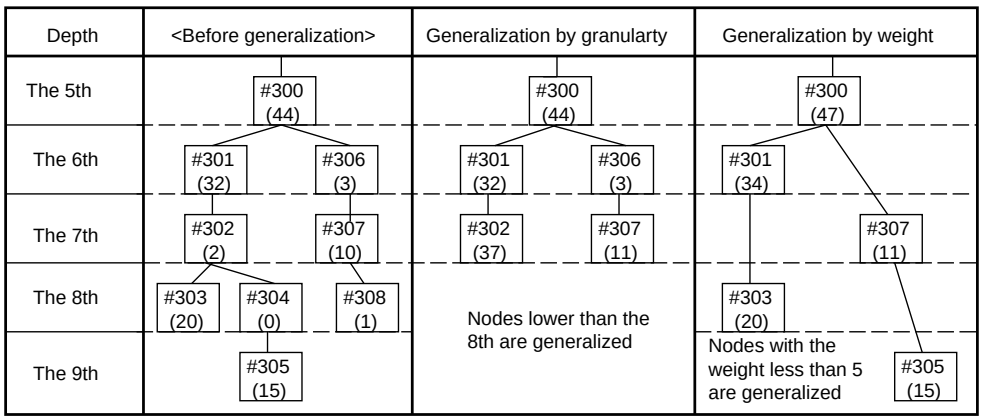

Figure 2: Generalization Methods

3.2. Minimum set of Vector Bases

In VSM, performance of information retrieval will decrease with the decrease in the dimension. Now, let us consider how to find the minimum set of vector elements that will minimally decrease the performance.

(1) Generalization by Granularity

The granularity of the meanings represented by a literal word is smaller than that of a semantic attribute. Accordingly, the granularity of a Semantic-Vector is rough compared to a Word-Vector. And in accordance with the generalization by granularity of semantic attributes, the performance of information retrieval will decrease. Then, while generalizing the bases, the change in performance is traced to find the minimum set of bases.

(2) Generalization by Weight

In this case, the target of generalization will be the semantic attributes not frequently used in the documents. But such a semantic attribute does not always become the target. Here, assuming that all documents in the database have the same probability to be relevant, let the summation of all document vectors be V_t. When the semantic attributes $\#i$ have a small value in V_t, they have little influence on the total performance of information retrieval. The most appropriate system will be obtained when all of the attributes in V_t have a balanced weight.

Consequently, the semantic attributes that increase the weight imbalance of the elements of vector V_t should not be generalized even if they have small weight. Taking these conditions

into consideration, we show how to select the attributes to be generalized.

Now, let us define the specific vector V_t for all of the documents in the database as follows:

$$V_t = (n_1, n_2, \cdots n_i, \cdots n_m) \tag{4}$$

Here, n_i represents the total frequency of words in the database, the meaning of which is $\#i$. And m is the number of attributes used by a specific vector.

Let us introduce the evaluation function H to assess the weight balance of bases by their "*variation*".

$$H = (n_1 - \bar{n})^2 + (n_2 - \bar{n})^2 + \cdots + (n_i - \bar{n})^2 + \cdots + (n_m - \bar{n})^2 \tag{5}$$

Here, $\bar{n}$ represent the mean value of n_i.

$$\bar{n} = \sum_{i=1}^{m} n_i/m \tag{6}$$

According to the above discussion, generalization should be performed by selecting the semantic attributes $\#i$ which decrease the value of H.

Now, let us consider the case in which a semantic attribute $\#i$ is generalized into the upper node $\#j$. n_i is added to n_j and m decreases by 1. Let the evaluation function be H' after the generalization. The change of the evaluation function $\Delta H(= H - H')$ is given as follows:

$$H - H' = (n_i - \bar{n})^2 + (n_j - \bar{n})^2 - (n_i + n_j - \bar{n})^2 \tag{7}$$

Letting $H - H' > 0$ as a condition, we obtain the following relation:

$$n_i n_j < \bar{n}^2/2 \tag{8}$$

From this relation, we find that an attribute $\#i$ that satisfies the condition (8) should be generalized.

Thus the generalization procedure is as follows:

① Generalize the semantic attribute $\#i$ where $n_i \cdot n_j$ is smallest.
② Experimentally evaluate the performance of information retrieval. If the degradation exceeds the threshold, stop or else return to ①.

3.3. Generalization Cost

Here, let's consider the calculation cost for generalization. Let the number of documents in the database and the bases of the specific vector be N and k respectively. It is known that the calculation cost is proportional to $(N + k)^4$ or $(N + k)^5$ for Word-VSM. On the other hand, it is proportional to $(N + k)^2 * k^3$ for LSI. Therefore, conventional methods are very expensive for large databases.

To compare to these, let M and d be the number of the attributes used for a specific vector and the maximum depth of the semantic attribute system, respectively. In our case of "*A-Japanese Lexicon*", $M = 2,710$ and $d = 12$. The generalization cost is proportional to $M \cdot d$ and $M^2 - k^2$ for generalization by granularity and generalization by weight, respectively.

4. EXPERIMENTAL RESULTS

4.1. Experimental Conditions

(1) Documents for Experiments

Here we use the test collection **BMIR-2** (Kitani 1998) registered to TREC for information retrieval. This collection is composed of documents, queries and their results. The documents comprise 5,079 articles from "*Mainichi Newspaper*" in the fields of economics or engineering. The queries have a unique format of "Retrieve articles on X," where several noun phrases are given for the variable X.

(2) Evaluation Parameters

Let us define three evaluation parameters such as R (recall factor), P (precision factor) and F (F-parameter) as follows:

$$R = \frac{No.\,of\ relevant\ documents\ retrieved}{No.\,of\ relevant\ documents\ in\ DB} \tag{9}$$

$$\tag{10}$$

$$P = \frac{No.\,of\ relevant\ documents\ retrieved}{No.\,of\ documents\ retrieved} \tag{11}$$

$$\tag{12}$$

$$F = \frac{(b^2+1)\cdot P \cdot R}{b^2 \cdot P + R} \tag{13}$$

Here, parameter b in *eq.* (11) represents the relative weights of R and P. We set $b = 1$ in our experiments.

4.2. Performance of IR

The experiments of information retrieval were conducted using 2,710 semantic attributes for bases of specific vector and the results were compared to conventional W-VSM.

4.3. (1) Experimental Procedure

One of the documents is selected for the query sentences from the articles which belong to the same subject. The examination was conducted to find the other articles from 5,079 documents. Changing the query sentences, this procedure was repeated 90 times and average performances were evaluated.

In order to compare the results to the conventional method, experiments for W-VSM were also performed. In this case, "$tf \cdot idf$" was used to determine the dimension and the nouns were used for bases in the order of the value of "$tf \cdot idf$".

4.4. (2) Experimental Results

The performance of information retrieval is shown in Fig.3 and, from this result, the value of F parameter was calculated as shown in Fig.4.

From these results, we can make the following observations:

(1) At any point of similarity, S-VSM yields higher recall performance but lower precision

compared to W-VSM.

(2) Maximum value of F is almost the same for both methods.

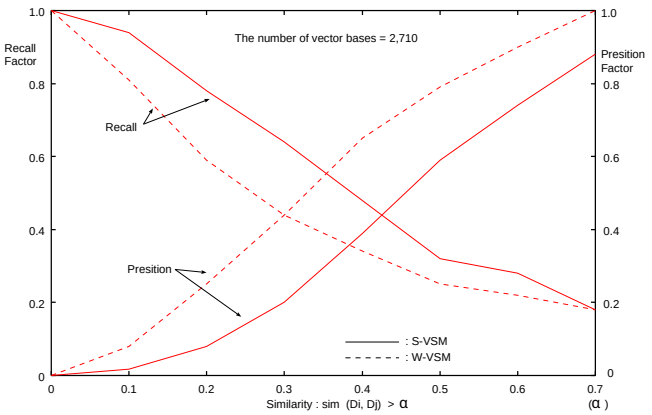

Figure 3: Similarity and Performance of Information Retrieval

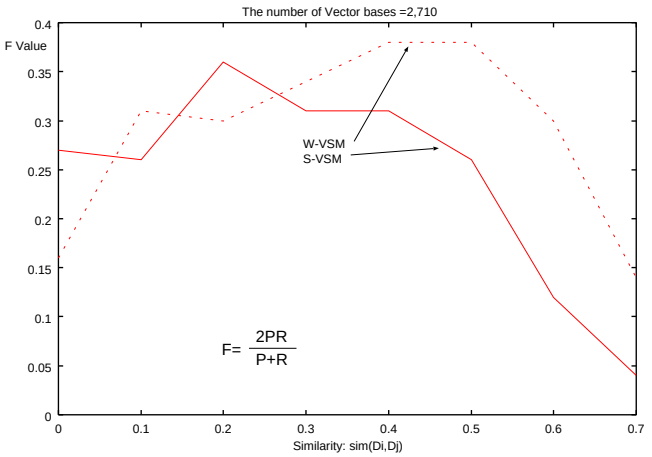

Figure 4: Similarity and F value

4.5. Reduction of Dimension

In order to reduce the number of vector bases, generalizations by granularity and weight were conducted. The relation between the number of bases and the performance of information retrieval is shown in Fig.5. The value of evaluation function H is also shown in the same figure.

From this figure, the minimum set of vector bases at which the performance of information retrieval does not decrease more than 10% or 20% from the maximum value was obtained as shown in Table 1.

From these figures and the table, the following observations can be made:

(1) S-VSM is robust in reducing the number of vector bases compared to W-VSM.

(2) In particular, generalization by weight is more robust than generalization by granularity.

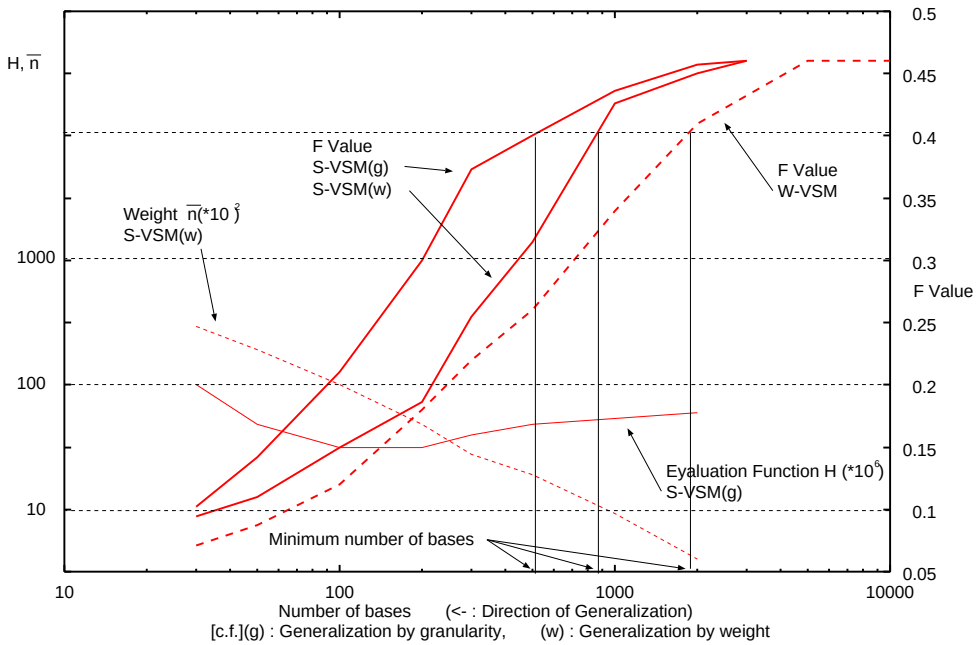

Figure 5: Determination of Minimum Number of Vector Bases

Table 1: Minimum Number of Vector Bases

Methods of VSM	Reduction of the vector bases	Permissible degradation of the performance	
		10% from the max	20% from the max
Propsed Method	Generalization by granularity	900 attributes	700 attributes
	Generalization by Weight	600 attributes	300 attributes
Conventional method	tf・idf	2,200 words	1,500 words

On condition that the performance of information retrieval does not decrease more than 10% to 20% from the maximum value, conventional W-VSM requires 2,000 dimensions. In comparison with this, the number of dimensions can be reduced to 300-600 in S-VSM.

5. CONCLUDING REMARKS

This paper has proposed Semantic-VSM, which uses the Semantic Attribute System defined by "*A-Japanese Lexicon*" as bases for specific vectors. Taking notice of the semantic relations between Semantic Attributes, generalization methods were also proposed in order to reduce the dimension without decreasing the performance of information retrieval.

In the experiments, this method was applied to the test collection of BMIR-J2, which contains 5079 newspaper articles. The results are as follows:

First, this method yields high recall performance compared to conventional Word-VSM because it is independent from fluctuations of word appearance, and documents which have synonyms are also retrieved. On the other hand, it is apt to pick up irrelevant documents and precision of performance decreases. Ultimately, the total performance (F value) is almost the same as W-VSM.

Second, in this method, generalization of vector bases can be easily performed at very low cost and the dimension can be greatly reduced compared to conventional W-VSM. Accordingly, the new method can be applied to large databases.

The remaining problems are as follows:

The first is related to how to select the target of generalization. In this study, the generalization target consisted of the nodes with small granularity or light weight. However, it can be pointed out that some of the large granularity nodes are apt to pick up irrelevant documents.

The second will be the problem of polysemy. This study did not consider the influence of polysemy, but the semantic attribute system used here has the ability to remove the ambiguity of meaning of words used in actual sentences.

These problems should be considered in the next step.

REFERENCES

(Borko and Bernick 1963): H. Borko, M. D. Bernick: Automatic Document Classification, Journal of the ACM, Vol.10, No.3, pp.151-162

(Burkley et al 1996) Burkley, Chris, A. Singhl, M. Mitra, G. Salton; New Retrieval Approaches using SMART; TREC4. In D. K. Harman (ed.),The second Text Retrieval Conference (TRC2), pp.25-48, 1996

(Deerwester et. al, 1990): S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Andauer, R. Harshman: Indexing by Latent Semantic Analysis, Journal of the Society for Information Science, Vo.41, No.6, pp.391-407

(Faloutsos and Lin 1995) Faloutsos, C. and Lin,K-I. : FastMap: A fast algorithm for indexing, datamining and visualization of traditional and multimedia data sets, Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data, pp.163-174

(Golub et al. 1965) G. H. Golub, W. Kahan; Calculating the Singular Values pseudouinverse of a matrix, SIAM J. Numer. Anal., 2 1965 205-224

(Golub and Loan 1996) Golub, G. H. and Van Loan,C. F. : Matrix Computations", The Johns Hopkins University Press, Third Edition

(Ikehara et. al, 1997) S. Ikehara, M. Miyazaki, S. Shirai, A./ Yokoo, K. Ogura, H. Nakaiwa, Y. Ooyama, Y. Hayashi: "A-Japanese Lexicon" Iwanami Bookstore

(Jardin et.al, 1971) Jardin, N. and van Rijsbergen, C.J.: The use of hierarchic clustering in information retrieval, Information Storage and Retrieval, 7, pp.217-240

(Jiang et al. 1999) M. W. Jiang,.J. M. Berry, J. M. Donato, G. Osrtouchov; Mining Consumer Product Data Via Latent Semantic Indexing, Intelligent Data Analysis 3:5, pp.377-398, 1999

(Kitani 1998) Kitani et.al,: Test Collection for Japanese IR: BMIR-J2, ISPJ Report, 98-DBS-114

(Kwok and Chan 1998) K. L. Kwok, M.Chan; Improving Two Stage ad-hoc retrieval for short queries, In SIGIR'98, pp.250-256, 1998

(Lin et al. 1998) S-H.Lin, C-S. Shih, M. C. Chen, J-M. Ho, M-T. Ko, Y-M. Huang; Extracting Classification Knowledge of Internet Documents with Mining Term Association; A Semantic Approach, Proc. of the 21st Annual Inernational ACM SIGIR Conference on Reaserach and Development in Information Retrieval, 1998

(Salton and McGill 1983) Salton, G. and McGill, J. M. (Eds.): Introduction to Modern Information Retrieval, McGill-Hill