Proceedings of NTCIR-4, Tokyo, April 2003 - June 2004

Patent Map Generation Using Concept-Based Vector Space Model

Hideyuki Uchida Atsushi Mano
BearNet Inc.
3-10-8 Tamagawa, Setagaya-ku, Tokyo 158-0094, Japan
rd-bearnet@basic.ne.jp

Takashi Yukawa
Nagaoka University of Technology
1603-1 Kamitomioka-cho, Nagaoka-shi, Niigata 940-2188, Japan
yukawa@vos.nagaokuat.ac.jp

Abstract

This paper proposes a patent map generation system using concept-based vector space model and presents evaluation results from the NTCIR-4 patent feasibility study (FS) task. The concept-base is a knowledge base of words, which expresses each word as an associated vector. The word vectors are computed based on word co-occurrence in a target document set. Therefore, the word vectors reflect target documents' characteristics. Each document in the target document set is expressed as a vector that is composed of vectors associated with words included in the document. The word vectors and document vectors are positioned in an identical vector space and the relevant degree of similarity between any two words and/or documents can be computed as a cosine coefficient of the two vectors. Taking advantage of this model, problems sections and solutions sections of patent documents are expressed as vectors, then, they are clustered and the label word for each cluster is chosen from words which give high cosine coefficient to the center of gravity of the cluster. A trial of generating patent maps for NTCIR-4 patent FS task topics using the system has been done. Comparing with human-generated patent maps, the system provides fairly good accuracy of clustering of target patents but poor accuracy of cluster labeling.

Keywords: *Patent Map, Concept Base, Vector Space Model, Hierarchical Clustering*

1 Introduction

The automation of patent map generation, especially for the commercial use, is in great demand. Manual generation of the patent map is very costly and having a limited supply. In order to examine the potential of the automatic generation of patent map, a task of organizing the given patent into two-dimentional matrix is created as in the NTCIR-4 patent feasibility study.

Our team had challenged on this task as we have created a clustering system by exploiting the Concept-Based Vector Space Model. Hence, the problem and the potential of the system were tested through our experiment. This paper explains on the method which had been used in the system and its evaluation result, in order to reveal subjects of future improvement.

2 Background

2.1 NTCIR patent map task

As mentioned in the overview paper, the task is to orgaize given patents into two-dimentional matrix. Criteria for the horizontal and the vertical axes of the matrix are also given and can vary depending on the topic. Each row and column of the matrix have to be labeled.

It is considered that systems should process the following two jobs: clustering or classifying given patents according to the criterial for the horizontal and vertical axes to map the patents into two-dimentional matrix, and finding proper label for each row and colunm from patent documents.

2.2 Concept-based vector space model

Expressing documents and queries as vectors in a multi-dimensional space and calculating the relevance or similarity as a cosine coefficient between two centroid vectors is known as the Vector Space Model [2]. With a basic relevance discernment scheme exploiting the vector space model, a vector of a document is mapped on a hyper-space where each keyword in the set of documents that correspond to an axis, such that the values along the axes for the documents correspond to the TF×IDF values for the keywords comprised in the documents. Because the scheme assumes a vector space in which the keywords directly correspond to the axes, there is the problem that synonyms and/or co-occurrences of keywords are not considered.

Some improved methods of solving the above problem have been proposed. One is Latent Semantic Indexing (LSI) by Deerwester [1]. This method first counts the occurrences of keywords throughout the documents and then constructs a word frequency matrix. Second, it reduces the rank of the matrix using Singular Vector Decomposition (SVD) and makes the reduced-rank matrix be the documents vector space.

Another is a co-occurrence based thesaurus (concept base) by Schütze [3, 4]. This method obtains a keyword vector space based on word co-occurrences in close proximities in documents, while LSI creates a document vector space based on word frequencies throughout documents. The keywords that co-occur in a similar manner throughout the documents are expected to be placed close to each other in the hyperspace. The vector for a document is represented as the center of gravity with keyword vectors comprised from it. Both methods are similar to each other in that a document vector is derived from a weighted average of vectors for keywords comprised in the document. In this method, documents having similar contents provide strong relevance even though the documents are not comprised of the same expressions. This differs from methods based on word occurrences, or boolean full-text search, in that a high relevance degree is obtained only when documents are comprised of similar expressions. We call this “concept-based vector space model.”

It should be pointed out for concept-based vector space model that a word and a document, which are different in nature from each other, are mapped together in the same multi-dimensional space. This means that the methods provide not only relevance between keywords, but also relevance between a keyword and a document, and between two documents.

2.3 Concept base construction

The concept base is a knowledge base of words, which is comprised of a set of words and their associated vectors. Each word is associated with a high dimensional vector (a word vector), and the vector is statistically calculated from the target document set. Figure 1 illustrates the construction procedure of the concept base. The procedure takes the following steps:

1. List every word that appears in the target documents. Let N be the number of words and w_i be i-th word in the word list.

2. Create $N \times N$ zero matrix. Let $\mathbf{C}$ be the matrix and c_{ij} be a i-th row and j-th column element in $\mathbf{C}$.

3. Count the co-occurrence of words throughout the documents: if word w_i and word w_j co-occur within the specific distance in a sentence, increment c_{ij}.

4. Reduce the rank of $\mathbf{C}$ to M using SVD, then obtain reduced-rank matrix $\mathbf{C}'$ (Nrows × Mcolumns).

5. $\mathbf{C}'$ forms the concept base. i-th row of $\mathbf{C}'$ corresponds to the word vector for word w_i.

Due to computing resource limitations, N cannot exceed 10,000. Thus, the word list is truncated based on occurrence count after step. 1. Though M can be 1 to N in principle, we use $M = 100$ because it is reported that this value is appropriate to discern similarity between words [4].

2.4 Clustering algorithms

There are two types of clustering algorithms: one is k-clustering which gives a partition of data points into k subset where k is fixed integer, the other is hierarchcal clustring which produces a hierarchy in which nodes represent subsets of data points simulating the structure found in the date set. Heirarchical clustering is supposed to be appropriate for patent map generation using vector space model. Because the number of clusters cannot be determined prior to start clustering.

Single linkage or Ward’s algorithm [5] is known as the most common a hierarchical clustering algorithm. Assuming S be a set of date points and n be the number of data points in S, the algorithm produces hierarchy with the following steps:

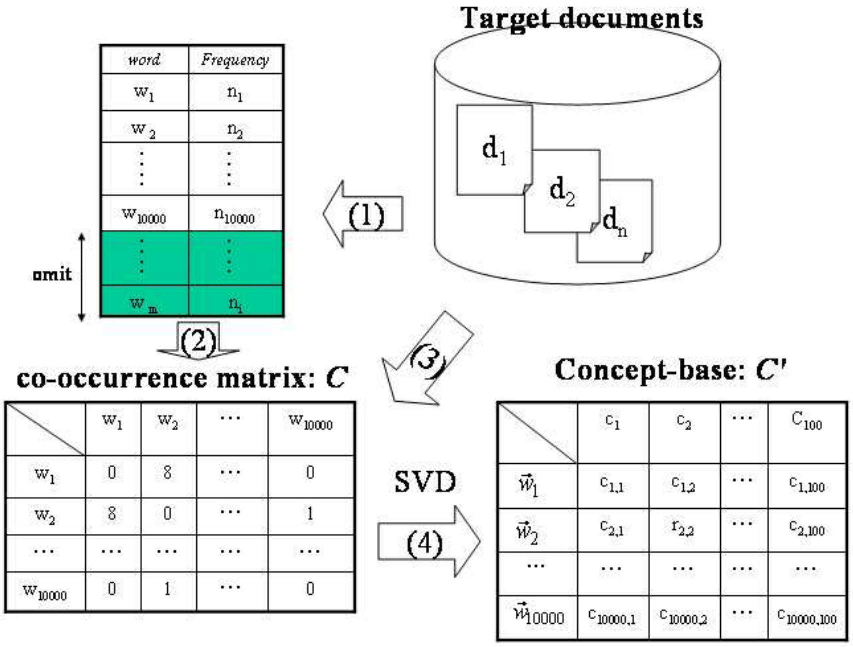
1. Place each instance of S in its own cluster (singleton). Note them as $S_1, S_2, S_3, ..., S_{n-1}, S_n$.

2. Compute a distance between every pair of elements in L and find the two closest clusters $\{S_i, S_j\}$.

3. Merge S_i and S_j to create a new internal node S_{ij} which will be the parent of S_i and S_j.

4. Go to step 2 until there is only one set remaining.

This is very basic algorithm for hierarchical clustering and its complexity is $O(N^3)$. Several algorithms have been proposed to reduce complexity. However, for the patent map generation task, the complexity of the basic algorithm does not lead severe problem because data set is relativly small (the number of patent is less than 100 for each topic).



word	Frequency
w_1	n_1
w_2	n_2
⋮	⋮
w_{10000}	n_{10000}
⋮	⋮
w_m	n_i

	w_1	w_2	⋯	w_{10000}
w_1	0	8	⋯	0
w_2	8	0	⋯	1
⋯	⋯	⋯	⋯	⋯
w_{10000}	0	1	⋯	0

	c_1	c_2	⋯	c_{100}
$\vec{w}_1$	$c_{1,1}$	$c_{1,2}$	⋯	$c_{1,100}$
$\vec{w}_2$	$c_{2,1}$	$r_{2,2}$	⋯	$c_{2,100}$
⋯	⋯	⋯	⋯	⋯
$\vec{w}_{10000}$	$c_{10000,1}$	$c_{10000,2}$	⋯	$c_{10000,100}$

Figure 1. Constructing Procedure of the Concept Base

3 A patent map generation system using concept based vector space model

3.1 A method of generating patent map

As descibed in the previous section, the vector space of a concept base reflects relations between words in the context of target document sentences. It is possible to say that aword vector in the concept base implies ontological information and fragment of knowledge on the target technology area. Thus a document vector, which is a normalized summation of the vectors for the words included in the document, can be seen as an aggregation of the concepts comprising the document and suggests the siginificance of it. Taking advantage this point, the patent map generation would be achived by clustering patent documents into categories according to degree of similarity of their document vector.

Since concept based vector space model locates word vectors and document vectors together in an idential vector space, retrival can be bilaterl, which means that the system can find not only relevant document fo a set of query words but also relevant word for a set of query document. The system can determine the label of cluster or at least provide the candidate words of label with retriving words which have high similaity degree with the center of gravity of the cluster.

The definitive pocedure of our patent map generation method is as follows:

Firstly, the sentences that coincide with mapping criteria are extracted from patent documents. The concept-base, which consist of vectors for every word used in portions is generated. A vector for patent document is calculated as the summation of vectors the words consisted in the concept-base. Then, the degree of similarity among patent document is determined. From these result, patent documents which are expressed as a multidimentional vector can be classified to categories referring to their degree of similarity by using the hierarchical clustering algorithm.

List of cluster's label candidate for all categories is obtained by calculating the cosine coefficient between the word vector and the value of center of gravity of the category. Fundamentally, the cluster's label should be represent the content of it's group. However, when we tried to fully depend on the degree of similarity results to choose the cluster's label, the problem of the irrelevant cluster's name occured. As we are still unable to solve those problem, cluster's name are determined manually by creating a nominal phrase from the top ranked of cluster's label candidate.

3.2 Implementation

We have implemented a prototype system exploiting the method as illustradted in Figure 2.

This system is composed by four main modules. The first module is the module for the extraction of target documents which consist of "problem to be solved" section or "solutions" section in patent summary from the patent document collection. The second one is the module to generate the concept base and document vectors. The third module is the clus-

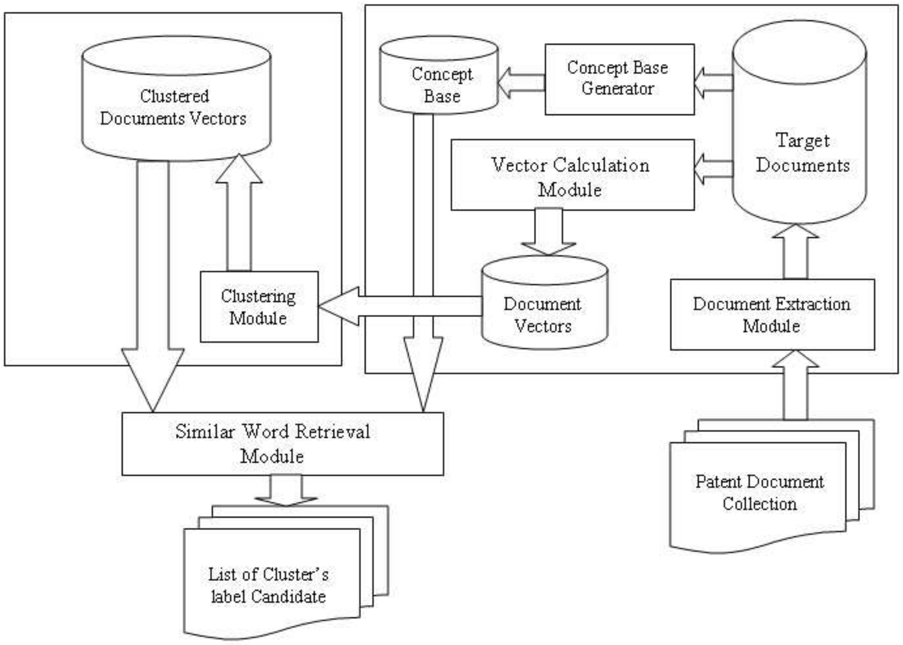


Figure 2. System architecture

tering module which classifies the target document's vectors into several groups. In the forth module, the similarity calculation module, the similarity degree of word vectors and each vector of the center of gravity for the cluster are computed to generate the candidate of the cluster's label.

4 Performance evaluation

In this section, the results which we have been obtained for the NTCIR-4 Patent fs task and thier evaluation by the human expert are shown.

The organizers of the task provided five topics. For each topic, the relevant patent document and criteria for the horizontal and vertical axes of the matrix are given. We use only the summary part of a patent document that is written in Japanese, as the main part of the document consist of too many irrelevant words to the topic, that will cause the document vector to be distorted. Though the relevant patents for each topic are given by the organizer, the patents which neither included "problem to be solved" or "solutions" section in the summary are omitted. Topics are as follws:

- Topic12. Composed from the patent documents that are related to the "blue light-emitting diode".The number of patent documents is 97.

- Topic24. Composed from the patent documents that are related to "solid high-polymer-type fuel". The number of patent documents is 98

- Topic25 is composed from the patent documents that are related to "Ultra hydrophilization of plastic surfaces". The number of patent documents is 99

For these topics, x axis stand for "problem to be solved " and y axis is "solutions". On the other hand, specific criteria are given for topics 7 and 8 as shown below:

- Topic8. Composed from the patent documents that are related to the "Hair Care Cosmetic Products". For this topic, x axis stand for "form of product" and y axis is "date of publication".The number of patent documents is 32.

- Topic7. Composed from the patent documents that are related to the "Gasoline-direct-injection engine". For this topic , x axis is "expression the concave" y axis is "piston top face".The number of patent documents is 62.

For topic 8, unlike in the case of topic12, 24 and 25, we use concatenation of "problem to be solved" section and "solutions" section for clustering along x axis. For y axis, we assume patents which have same "date of publication" belong to the same cluster.However for this topic, patent has difference date of publication, therefor each patent belongs to its own cluster.

For topic 7, most of the words exist in the top list of cluster's label candidate were unreleted word. Therefore, we gave up on getting the result.

Table 1 shows the result of these topics compared with the resultevaluated by the expert.Table 2 shows

Table 1. Evaluation result of topic 12, 24 and 25

	○	×	⊗
Topic12	76.5	13.3	10.2
Topic24	80.8	7.1	12.1
Topic25	97.0	1.0	2.0

Table 2. Evaluation result of topic 8

	○	×	⊗
Topic8	27.2	57.6	15.2

the result of topic8 results which evaluated by the expert.

○ represents in table represents in the percentage of the system-generated clusters which coincided with those evaluated the human expert, × represents the percentage of the system generated cluster which differed from those evaluated by the human experts, ⊗ represents the percentage of documents omitted in our system.

5 Discussion

The system that was created to classify the patent document collections which are provided by NTCIR-4 patent fs task and its evaluation results were explained. Topics(patent document collections) was catogarized into two groups:

1. A group with the general mapping criteria; x axes stand for the "problem to be solved" and y axces stand for "solutions", as the related section for each axes were contained in the patent summary(topic 12, 24, and 25).

2. A group with the specific mapping criteria; the criterion is often a deep technical aspect of the taget area(topic 7 and 8).

For the former group, our system comparatively made a good classification, though most of the labels for the categories differ from those provided by the human experts. For the later group, both the result of classifying patent documents into categories and labeling were too poor.

For each groups, A the reason of failure labeling method is as follows: Most of category's name that is provided by the human experts is a compound word. However, list of clueter's label candidate that is generated by our system does not included compound words due to difficulties of determining them. Therefore this phenomenon was happend. If the compound words are expressed as the mutlidimentional vector, it may be possible to improve this method.

For the later group, as the reason failure to classify the patent documents into categories is as follows: The senetences extracted from patent documents are mostly include few words that are related to criteria for each axes. Therefore, the system fails to classify them into proper categories. In fact, most of the word related to the mapping criteria exist in the claims section more than the summary section. For that reason, in order to ensure the system functions properly, the claims section should be used for the classification too.

6 Conclusion

We proposed the method for generating patent map automatically, implemented it, and examined it's potential according to experiment with the use of patent documents provided by NTCIR4 fs task.

The system generated patent maps are compared with those constraced by the human experts. In the result, the system made a relatively good classification of the documents into categories according to their characterlistics.

The result suggests that automated patent map generation is not unrealistic in despite of its complexity. Though, the proposed method is rather naive because the research focuses on investigating its feasibility and clarifying issues.

Extracting portions including informative passeges from the patent document for classifying and handling the compound words appropriately in the concept base are necessary to improve its accuracy. These enhancements are as future works.

References

[1] S. Deerwester, S. T. Dumais, G. W. Furnas, et al. Indexing by latent semantic analysis. *Journal of American Society for Information Science*, 41(6):391–407, 1990.

[2] G. Salton and C. Buckley. Team weighting approaches in automatic text retrieval. In K. S. Jones and P. Willet, editors, *Readings in Information Retrieval*, pages 323–328. Morgan Kaufmann Publishers, 1998.

[3] H. Schütze and J. O. Pedersen. A cooccurrence-based thesaurus and two applications to information retrieval. In *Proc. RIAO '94*, 1994.

[4] H. Schütze and J. O. Pedersen. Information retrieval based on word sense. In *Proc. 4th Annual Symposium on Document Analysis and Information Retrieval*, pages 161–176, 1995.

[5] J. H. Ward, Jr. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58:236–244, 1963.