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 from vectors associated with words included in the document. The word vectors and document vectors are positioned in an identical vector space and relevant degree between any two words and/or documents can be computed as a cosine coefficient of 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 are chosen from words which gives 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 corresponds 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 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 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. In more detail, the concept base is constructed with the following steps:

1. List every words appeared 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 are 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 $\times$ 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, 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 cluster cannot be determined prior to start clustering.

Single linkage or Ward's algorithm [5] is known as the most common 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 are relativly small (the number of patent is less than 100 for each topic).

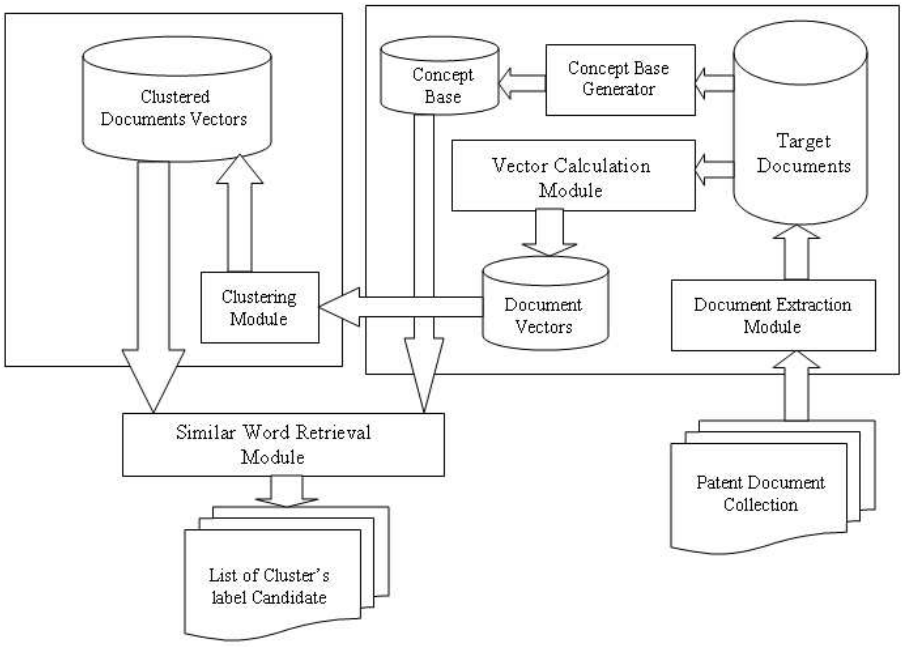

Figure 1. System architecture

3 Implementation

In this section, the architecture of the system used in our resarch in order to automate the classification of particular document set and its process flow are introduced. Additionaly, The manual method for labeling the clusters which was automatically classfied by the system is discribed.

3.1 System architecture and process flow

This system is composed by four main modules. The first module is the module for the extraction of target documents which are 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 clustering module which classifying the target document's vectors to several groups, and 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. The architecuture of this system is illustrated in Figure 1.

We introduce process flow of the system. The system firstly maps each patents with clustering section of the patent document, the produces the list of cluster's label candidates. Datialed procedure for this process is as follows:

1. For each axis, the corresponding sections in patent documents are extracted. These sections are treated as the target documents in our concept-based clustering system.

2. The concept-bace is constructed from the target documents with the procedure described in subsection 2.3. Then, the document vector for each target document are calculated based on the concept base.

3. For each axis, the hierachial clustering is applied to the vectors (document vectors) which correspounds with an axis and the vectors are classfied to several document clusters.

4. For each axcis, the similarity degree of each word vector and each vector of the center of gravity for the cluster are computed and the obtained label are given list of cluster's label candidate.

3.2 Labeling method

As one of the assignment in the patent task, the patent document collection needs to be classified, and then the classified clusters need to be labeled accordingly. For that reason one of the important parts is to consider the method of labeling the clusters.

Fundamentally, the cluster's labels should represent the content of its group. When we tried to fully depend on the system to choose the cluster's label from

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

the words with largest similarity degree in the list produced by the system, the possibility of the generation of same label for different cluster was occurred. The other problem was the irrelevant cluster's name problem.

In order to avoid these problems cluster's labels are determined manually by selecting some words from the top ranked words of the list of cluster's label candidate and concatenate it as a nominal phrase.

4 Performance evaluation and discussion

In this section, the results which we have submitted and the result evaluated by the expert are compared and examined for each topic. For each topic, the relevant patent document and criteria for the horizontal and vertical axes of the matrix are given. We use only summary part of a patent document, as the main part of the document is consisting too many irrelevant words to the topic, that will cause the document vector 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.

Topic 12 is composed from the patent documents that are related to the "blue light-emitting diode". Topic 24 is composed from the patent documents that are related to "solid high-polymer-type fuel". Topic25 is composed from the patent documents that are related to "Ultra hydrophilization of plastic surfaces". For these topics, x axes stand for "problem to be solved " and y axes is "solutions". Table 1 shows the result of these topics compared with the resultevaluated by the expert.

◯ in table represents in the percentage of the system-generated clusters which coincided with the clusters constructed by the human expert (answerset), × represents the percentage of the system generated cluster which differed from the answer-set, ⊗ represents the percentage of documents omitted in our system.

For the topic 12, 24 and 25 as shown in the table1, the system had comparatively made a good classification. However, most of labels for the clusters in the submitted result differ from those provided by the human experts. Moreover, Human expert indicated that the number of clusters generated by the system is too many.

Table 2. Evaluation result of topic 8

	◯	×	⊗
Topic8	27.2	57.6	15.2

Topic8 is composed from the patent documents that are related to the "Hair Care Cosmetic Products". For this topic, x axes stand for "form of product" and y axes is "date of publication". Unlike in the case of previous topics, we use concatenation of "problem to be solved" section and "solutions" section for clustering along x axis. For y axis, we assume patents which having 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.Table 2 shows the result of topic8 results which evaluated by the expert.

For topic8, as shown in table2, although our system were able to map the given patents and provide the lists of label candidate which includes few word relevant to the clusters, most of results did not coincide with the answerset. Topic7 is composed from the patent documents that are related to the "Gasoline direct-injection engine". For this topic , x axes is "expression the concave" y axes is "piston top face". For this topic, most of the words exist in the top list of cluster's label candidate were unrelated word . Therefore, we gave up to submit this result.

5 Conclusions and future works

The system that was created to classify the patent document collection which was provided by NTCIR-4 patent FS task and its evaluation results was explained. Topics was catogarized into two groups:

1. Group with the obvious 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.

2. Group with the un-obvious portion of mapping criteria.

For the former group, our system comparatively able to categorized the patent into clusters for all topic, though the labeling for clusters was not working properly. It is one of our future works to create a fully automated system which is able to label the clusters accurately.

For the later topic group, neither the result of categorizing patent into cluster nor labeling the cluster was bad. In order to ensure the system works properly for this type of group, the method of retrieving related section of the mapping criteria for each axes from the

patent document is indispensable. Overall, the method to realize these subjects is also still our future works to be examined.

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.