An Automated Research Paper Classification Method for the IPC system with the Concept Base

Takanori Shimano
Nagaoka University of Technology
1603-1 Kamitomioka-cho, Nagaoka-shi, Niigata 940-2188, Japan

Takashi Yukawa
Nagaoka University of Technology
1603-1 Kamitomioka-cho, Nagaoka-shi, Niigata 940-2188, Japan

Abstract

In the present paper, a classification method using the Concept Base is proposed and evaluated in the Patent Mining Task of the NTCIR-7 workshop. In this task, research papers are classified into the International Patent Classification (IPC) system. The classification enables research papers to be located on a patent map. In order to classify a paper, patent documents that are similar to the paper are retrieved, and the paper is classified as the class of the patent documents. The proposed approach can classify a research paper into correct classes when appropriate patent documents are retrieved. However, there is a problem in that patent documents differ from research papers with respect to document characteristics, even if the technical idea of a patent is the same as that of a research paper. For example, the terms and document structure used in a patent differ from those of a paper. Therefore, the classification method required for this task requires an approach that addresses this problem.

In order to clarify the performance of the naive method of this approach, the authors classified research papers using a simple classification method with the Vector Space Model (as the baseline method).

Then, in order to solve the problem caused by the term difference, a classification method using the Concept Based Vector Space Model (CBVSM) is proposed. The degree of similarity in this method reflects the semantics of the words.

The Mean Average Precision (MAP) value of the proposed method did not exceed that of the baseline method. The proposed method is classified in the worst group among the participating teams. However, the proposed method provides better average precision (AP) values than the baseline method for 33% of topics. Based on these results, it is suggested that a more effective method could be constructed using a combination of the baseline method and the method using CBVSM.

Keywords:*patent classification, IPC system, Concept-Based Vector Space Model*

1 Introduction

In the previous patent classification subtasks performed at the NTCIR workshop, patents were classified into an F-term category system. Past patents were classified into the system by hand. Therefore, the patents are used as training data for the classifier. Therefore, the classification method using machine learning is effective for that task.

In the NTCIR-7 patent mining task, research papers are classified into an IPC system. Since research papers do not have IPC codes, direct training data are not obtained. On the other hand, each patent document is classified into the IPC system. Therefore, patent documents are used as training data. A machine learning approach can be used for previous NTCIR tasks. However, this approach does not provide good performance because the document types of the target and the topics are different. An approach that retrieves patent documents using a target and that classifies the paper according to the class of the retrieved patent document would perform well. Therefore, in the present paper, the authors use this approach.

The approach requires the retrieval of the correct patent documents for correct classification. However, the retrieval encounters problems such as term differences and document structure differences, which degrades the accuracy of retrieval.

Therefore, the classification method using the Concept Based Vector Space Model (CBVSM) in retrieval is proposed. The retrieval method can retrieve documents based on semantics. The method is expected to solve the problem of term differences. A simple classification method using the Vector Space Model (VSM) is used as the baseline method for comparing the performance of the proposed methods. The proposed methods are evaluated with the test set of the NTCIR-7.

The results of the evaluation are also presented and discussed.

2 Background

This section presents a brief introduction to background knowledge on the patent mining task at the 7th NTCIR workshop and the document retrieval method using the CBVSM.

2.1 NTCIR-7 Patent Mining Task

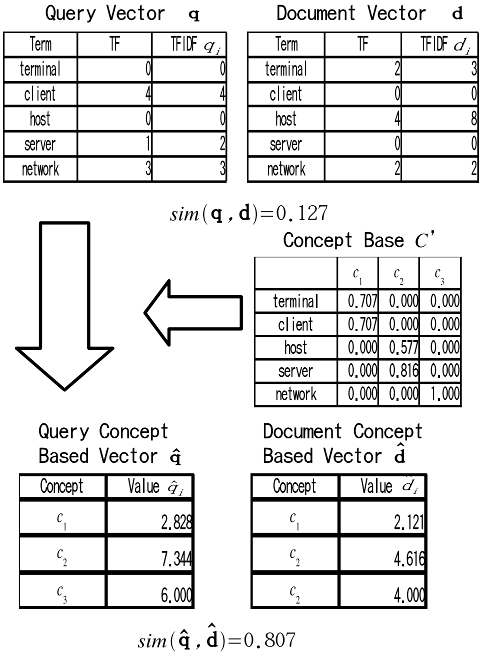
In the NTCIR-7 patent mining task [3], research papers are classified into the International Patent Classification (IPC) system. The IPC system is based on an international multi-lateral treaty administered by WIPO [1]. The IPC system is a standard patent classification system and is used worldwide.

In the NTCIR-7 task, over 3 million of Japanese patent documents published in 10 years from 1993 to 2002, approximately 740,000 research papers (titles and abstracts) are provided as the training data. A total of 897 topics are set as the target topics. The topics are structured as a title and an abstract of a research paper. Participants classify the topics using their own systems, and the classification results are then submitted to the task organizer. The organizer evaluates the results with A-precision (AP) and Mean Average Precision (MAP). A-precision is an index of accuracy for ranked multiple answers. In the evaluation process, AP values are calculated for the results of each target topic, and the MAP value, which is the mean of the AP values, is then obtained. Finally, the organizer releases the evaluation results to the task participants.

2.2 Concept Based Vector Space Model

The Concept Base [5] associates a multi-dimensional space with words in documents. Vectors of semantically similar words point in the same direction. Assuming a document as bag of words and expressing the document vector as the composition of the vectors of the words included in the document, the model is assumed to be a variation of the VSM. The degree of similarity is defined as a cosine coefficient of two document vectors. In the retrieval, documents are retrieved based on the degrees of similarity. In the CBVSM, vectors for documents that include no common words, but only semantically similar words, point in a similar direction. Therefore, the CBVSM can achieve information retrieval based on the semantics of documents.

Figure 1 shows an example of the document retrieval using the CBVSM. Document Q is used as query q. The query q includes words such as ”client”, ”server”, and ”network”. The document d is a retrieved document that includes words such as ”host”, ”terminal”, and ”network”. When the degree of similarity is calculated with words, the document vectors, ”client” and ”terminal” (or ”server” and ”host”) do not contribute to the degree of similarity because they are different terms. However, if ”client” and ”terminal” (or ”server” and ”host”) are synonyms, in the CBVSM, their vectors point in similar direction. In addition, the similarity between the query and the document calculated with the CBVSM becomes higher than that calculated using the simple VSM.



Query Vector $\mathbf{q}$

Term	TF	TFIDF q_i
terminal	0	0
client	4	4
host	0	0
server	1	2
network	3	3

Document Vector $\mathbf{d}$

Term	TF	TFIDF d_i
terminal	2	3
client	0	0
host	4	8
server	0	0
network	2	2

$sim(\mathbf{q}, \mathbf{d}) = 0.127$

Concept Base C'

	c_1	c_2	c_3
terminal	0.707	0.000	0.000
client	0.707	0.000	0.000
host	0.000	0.577	0.000
server	0.000	0.816	0.000
network	0.000	0.000	1.000

Query Concept Based Vector $\hat{\mathbf{q}}$

Concept	Value $\hat{q}_i$
c_1	2.828
c_2	7.344
c_3	6.000

Document Concept Based Vector $\hat{\mathbf{d}}$

Concept	Value $\hat{d}_i$
c_1	2.121
c_2	4.616
c_2	4.000

$sim(\hat{\mathbf{q}}, \hat{\mathbf{d}}) = 0.807$

Figure 1. Similarity Calculation with the Concept Base

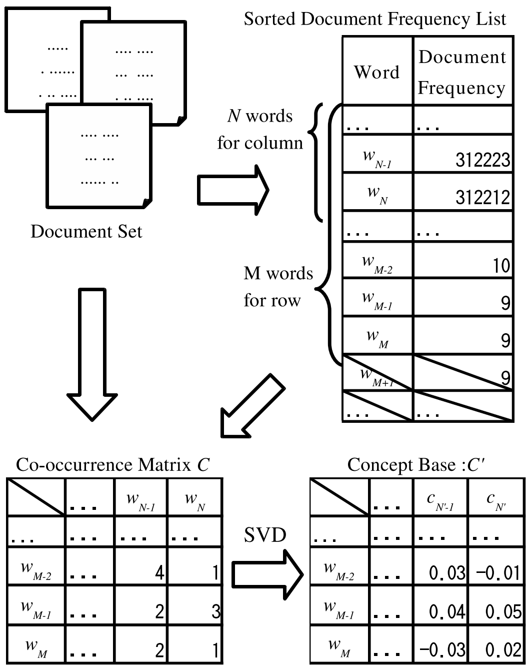


Figure 2. Construction of the Concept Base

2.3 Construction of the Concept Base

The Concept Base is a knowledge base of words that 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 a document set. The construction procedure of the concept base is illustrated in Figure 2. The procedure is made up of the following steps:

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

2. Create an $M \times N$ zero matrix. Let C be the matrix, and let c_{ij} be the i-th row and j-th column element in 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 C to N' by Singular Value Decomposition (SVD) and obtain reduced-rank matrix C' (M rows x N' columns).

5. C' forms the concept base. The i-th row of C' corresponds to the word vector for word w_i.

3 A Classification Method using the Simple Vector Space Model

In order to clarify the base performance of the proposed approach, a classification method using the simple VSM [4] is proposed as a baseline method. The proposed method classifies the target topics using the similarity between topics and classified patent documents.

When a target topic is given, the method retrieves patent documents that are similar to the topic with rank based on the VSM. The top 200 documents are then used for classification of the topic. Each of the retrieved patent documents has its own IPC codes. Therefore, the classes of the topics are decided with the IPC codes associated with the retrieved documents. In this section, the classification method using the VSM is described in detail.

3.1 Calculating Document Vectors

To calculate the degree of similarity of documents based on the VSM, a topic and patent documents are associated from text data into document vector space. Vectors of a patent document $\mathbf{d}$ and a topic $\mathbf{p}$ are weighted by the Term Frequency and Inverse Document Frequency (TF-IDF).

3.2 Classification

To classify a topic, patent documents similar to the topic are retrieved based on the VSM. The degree of similarity with the documents is calculated as the cosine coefficient between the vector of the topic and the vectors of the documents.

In the retrieval, the top 200 high-ranked patent documents are retrieved. The classes of the topics are decided from the IPC code associated with the retrieved patents. The classification score between a research topic and class C is calculated by the summation of the degree of similarity between the topic and the documents associated with the IPC code C.

$$score(\mathbf{p}, C) = \sum_{\mathbf{d} \in C} sim\,(\mathbf{p}, \mathbf{d}) = \sum_{\mathbf{d} \in C} \frac{\mathbf{p} \cdot \mathbf{d}}{|\mathbf{p}|\,|\mathbf{d}|} \qquad (1)$$

IPC codes are sorted by score and are listed.

4 A Classification Method using the Concept Based Vector Space Model

To solve the problem of term difference between topics and patents, a method using the CBVSM is proposed. The method associates the documents into the concept based vector space. Therefore, the patent documents are retrieved with synonyms in the classification. In this section, the proposed CBVSM-based patent classification method is described.

In the proposed method, the topics are classified in a manner similar to the method using the simple VSM. One difference is that the concept based vector of the documents and the topics is used at retrieval in the classification. A document vector of a patent document $\mathbf{d}$ is converted from a document vector into a concept based vector $\hat{\mathbf{d}}$ with the Concept Base $C^{'}$. A document vector of a topic $\mathbf{p}$ is also converted into a concept based vector $\hat{\mathbf{p}}$ by the Concept Base $C^{'}$. The concept base is constructed using the procedure described in Subsection 2.3.

The classification procedure is described in Subsection 3.2.

5 Evaluation at the NTCIR-7 Test Set

In this section, the evaluation results of the proposed methods for NTCIR-7 are described and compared with the results reported by all of the teams that participated in the task.

5.1 Evaluation results of the proposed methods

Upon a formal run, incorrect results of classification with systems that contained bugs are submitted. Therefore, the MAP in the overview paper (E1, E2 of Table 3) is too low. The reevaluated results of the systems in which the bugs have been fixed and reevaluated are shown in Table 1. The discussion of the present study is based on these results.

Table 1. Evaluation Results of Fixed Systems

Run ID	Mean average precision
nut1-1	0.2963
nut2-1	0.2388

5.2 Comparison with the Evaluation Results of NTCIR-7

Figure 3 shows the results of MAP values for all teams. A total of 24 systems of five teams participated in the task. The nut1-1-with-bug and the nut2-1-with-bug show the results for the system containing bugs, and nut1-1 and nut2-1 show the results for the fixed system.

5.3 Discussion

The MAP value for the CBVSM-based method was lower than the baseline method. However, in 33% of all topics, the AP value for the method is higher than that for the baseline method. Table 4 shows an example. In the baseline method, the degrees of similarity between the topic and the patents are calculated based on term-level matching. Therefore, when the terms of the document are not in the topic, their weights are not added to the degree of similarity, even if the concept of the terms is contained in the topic. On the other hand, the CBVSM-based classification method calculates the semantic similarity between the topic and the patent documents. If two different words that have similar meanings are contained in a topic document or a patent document, these words do not contribute to the degree of similarity in the baseline method, but they contribute in the CBVSM-based method. In case in which the precision of the method is higher than that of the baseline method, the topic is assumed to have these characteristics.

Table 2. Examples of topics in which the results of classification obtained by the proposed method are more accurate than those obtained by the baseline method

Topic-ID	Baseline method	Proposed method
302	0.1111	1.0000
315	0.1667	1.0000
336	0.2500	1.0000
339	0.2395	0.6276
440	0.4111	1.0000

The degree of similarity calculation of the topic #302 is described in detail as an example in which the classification precision of the CBVSM-based classification method is higher than the baseline method. Table 3 and Table 4 show the calculation of the degree of similarity using the simple VSM and the CBVSM, respectively. The document d_1 includes the same content as topic #302, and d_2 includes content that differs from that of topic #302. When d_1 is retrieved, the topic is classified with the correct IPC codes, but when d_2 is retrieved, the topic is classified with incorrect IPC codes. The word w_{21} included in the document d_1 is a synonym of the word w_7 included in the topic. In the baseline method, the words w_7 and w_{21} are different and so do not contribute to the degree of similarity. As a result, the degree of similarity of d_2 is higher than that of d_1. On the other hand, in similarity calculation in the CBVSM, the occurrence of words w_7 and w_{21} contributes to the concept c_{341} and approximates the vector of the correct patent to the vector of the topic. As a result, the degree of similarity to the patent increases and the correct IPC code is assigned to the topic. Therefore, it is suggested that the CBVSM-based method solves the problem of difference of terms.

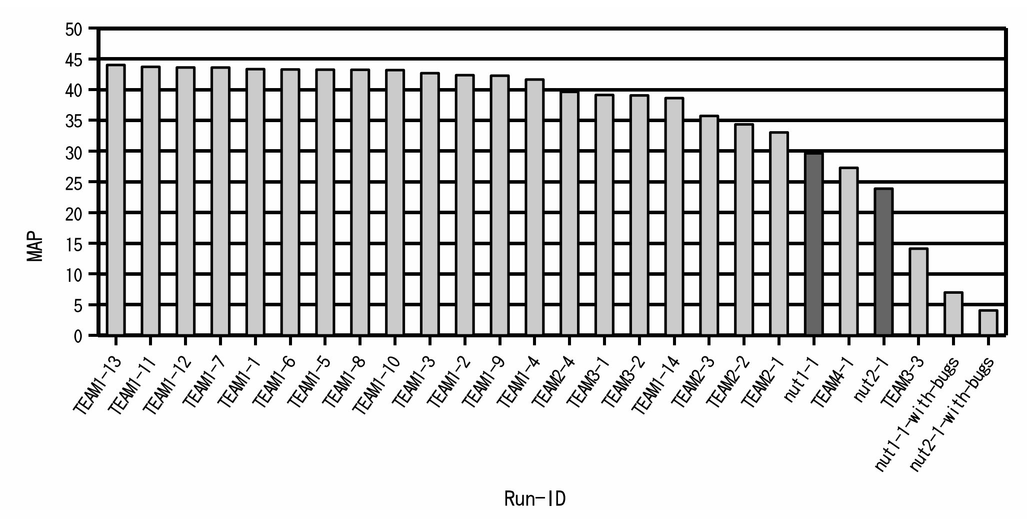


Figure 3. Evaluation Mean Average-precision of Results at NTCIR-7 and the Fixed System

Next, in both the baseline method and the CBVSM-based method, a problem was revealed whereby the result of the document retrieval is dominated by a number of insignificant terms. The documents are converted into vectors weighed with the TF-IDF. When a word having a large TF value is included in the document, if the IDF of the word is small, i.e., the word is insignificant, then the weight becomes large. The degree of similarity is influenced by these words, and the influence of the significant words is spoiled by them. For example, in the vectors shown in Table 3, the TF value of the word w_{10} (”polymerization”) is too large, and the word is weighted heavily in the vectors. However, the word is not useful for retrieval of patent documents in the topic. A method that sets a ceiling value is considered as a potential method for addressing this problem.

Table 3. Document Vectors of Topic #302 , Patent #95228636 and Patent #00313706

Word	#302 p	#95228636 d_1		#00313706 d_2	
	$\frac{p_i}{\lvert\mathbf{p}\rvert}$	$\frac{d_{1i}}{\lvert\mathbf{d_1}\rvert}$	$\frac{p_i d_{1i}}{\lvert\mathbf{p}\rvert\lvert\mathbf{d_1}\rvert}$	$\frac{d_{2i}}{\lvert\mathbf{d_2}\rvert}$	$\frac{p_i d_{2i}}{\lvert\mathbf{p}\rvert\lvert\mathbf{d_2}\rvert}$
...	...	...	...	...	...
w_7	0.1415	0.0895	0.0127	0.0277	0.0039
...	...	...	...	...	...
w_{10}	0.4772	0.5744	0.2741	0.7089	0.3383
...	...	...	...	...	...
w_{21}	0.0000	0.1600	0.0000	0.0000	0.0000
...	...	...	...	...	...
$sim(\mathbf{p}, \mathbf{d}_j)$			0.4760		0.5085

6 Conclusion

In the present paper, in order to establish automated research paper classification, patent documents that have similar technical content as the topic were retrieved and the IPC code of the topic with the code associated with the retrieved patent was decided. The problem of term difference is described and a method by which to solve this problem using the CBVSM is proposed.

A simple classification method using the VSM and a CBVSM-based classification method are implemented. The proposed method is evaluated using the data set of NTCIR-7. As a result of the evaluation, the MAP value of the CBVSM-based classification method was found to be less than that of the baseline method. However, in 33% of the topics, the AP value

Table 4. Concept-based Vectors of Topic #302 , Patent #95228636 and Patent #00313706

Concept	#302 p	#95228636 d_1		#00313706 d_2	
	$\frac{\hat{p}_i}{\lvert\hat{\mathbf{p}}\rvert}$	$\frac{\hat{d}_{1i}}{\lvert\hat{\mathbf{d}}_1\rvert}$	$\frac{\hat{p}_i \hat{d}_{1i}}{\lvert\hat{\mathbf{p}}\rvert\lvert\hat{\mathbf{d}}_1\rvert}$	$\frac{\hat{d}_{2i}}{\lvert\hat{\mathbf{d}}_2\rvert}$	$\frac{\hat{p}_i \hat{d}_{2i}}{\lvert\hat{\mathbf{p}}\rvert\lvert\hat{\mathbf{d}}_2\rvert}$
...	...	...	...	...	...
c_{341}	-0.2516	-0.1826	0.0459	-0.1070	0.0269
...	...	...	...	...	...
c_{346}	0.0867	0.0867	0.0075	0.1272	0.0110
...	...	...	...	...	...
c_{359}	-0.1489	-0.2142	0.0319	-0.1568	0.0233
...	...	...	...	...	...
$sim(\hat{\mathbf{p}}, \hat{\mathbf{d}}_j)$			0.8393		0.8367

of the classification obtained using the CBVSM-based classification method is higher than that obtained using the baseline method. In these topics, a word included in an appropriate patent document is not included in the topic document. However, the topic document contains a semantically similar word. Therefore, the CBVSM-based method is proposed in order to resolve the problem of term difference.

In the future, we intend to investigate two areas of concern: terms with large TF values and improvement of the CBVSM-based classification method for topics that are not classified well with the method. We will attempt to address the problem of a large TF value by setting a ceiling value. In addition, we will attempt to improve the accuracy of the classification using a combination of the CBVSM and the simple VSM.

References

[1] World Intellectual Property Organization http://www.wipo.int/portal/index.html.en

[2] Atsushi Fujii, Makoto Iwayama, Noriko Kando: Overview of Classification Subtask at NTCIR-6 Patent Retrieval Task, Proceedings of the 6th NTCIR workshop, 2007.

[3] Hidetsugu Nanba, Atsushi Fujii, Makoto Iwayama, Taiichi Hashimoto:
Overview of the Patent Mining Task at the NTCIR-7 Workshop, Proceedings of the 7th NTCIR workshop.
Proceedings of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access, 2008.

[4] Saltion, G., and Buckley, C. Term-weighting approaches in automatic text retrieval. Information Processing and Management 24 (1998), 513-523.

[5] Schütze, H., and Pedersen, J. O. 1994. A cooccurrence-based thesaurus and two documents to information retreival. Proceedings of RIAO '94.

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