Term Weighting Classification System Using the Chi-square Statistic for the Classification Subtask at NTCIR-6 Patent Retrieval Task

Kotaro Hashimoto[+]

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
yukawa@vos.nagaokaut.ac.jp

Abstract

In the present paper, a term weighting classification method using the chi-square statistic is proposed and evaluated in the classification subtask at NTCIR-6 patent retrieval task. In this task, large numbers of patent applications are classified into F-term categories. Therefore, a patent classification system requires high classification speed, as well as high classification accuracy.

The chi-square statistic can calculate the frequency of word appearance in the F-term and the frequency of word non-appearance in the F-term. The proposed method treats words as a scalar value and a ranking algorithm simply adds the word values of each word included in the test patent document in each F-term. Therefore, the proposed method provides classification that is significantly faster than other methods.

The proposed method is evaluated in A-precision, R-precision, and F-measure. Although the proposed method did not obtain the best score, this method achieves a classification accuracy that is as high as those of other methods using machine learning or the vector classification method.

In this task, the processing speed is not evaluated. Therefore, processing speed is also evaluated. The evaluation results show that the proposed method is much faster than that using the vector classification method.

Evaluation results of classification accuracy and processing speed show that the proposed method is confirmed to be effective and to be practical.

Keywords: *patent classification, F-term, chi-Square statistic*

1. Introduction

In the previous NTCIR workshop, a machine learning method and a vector classification method, such as the k-Nearest Neighbor method, provided good results for the classification subtask. However, these methods are expected to require a long processing time when classifying a large number of patent documents.

At the classification subtask, the number of F-term themes and test documents increase considerably. Therefore, the patent classification system is required to have a high classification speed as well as high classification accuracy.

In the present paper, a high-speed classification method having a classification accuracy that is as good as or better than those of traditional methods is proposed using a complex machine learning classification method. In addition, the details of proposed method are described herein. The proposed method is implemented and evaluated for a classified test collection in the classification subtask. The results of the evaluation are presented and discussed.

+Current affiliation: Nippon Telegraph and Telephone West Corporation

2. Background

2.1 Classification subtask at NTCIR-6

In the classification subtask at NTCIR-5 [2], two subtasks, the theme classification subtask and the F-term classification subtask, are performed. The F-term classification subtask is used for only five themes, including 2,562 test documents. These documents assign multiple F-terms.

In this NTCIR-5 F-term classification subtask, popular methods include the use of the K Nearest Neighbor method [3] and the Support Vector Machine method [4]. These methods provided good classification accuracy but require a long processing time.

2.2 Classification subtask at NTCIR-6

In classification subtask at NTCIR-6 [5] only the F-term classification task is performed. The total number of F-term themes is 108 and total number of test documents is 21,606, and these increase at a great rate for NTCIR-5. Therefore, the classification system requires faster classification. Although the Japan Patent Office (JPO) categorizes approximately 1,900 F-term themes and JPO has over 4.5 million patent applications, the NTCIR-6 test collection has a smaller number of documents, for the purpose of practical use.

In the evaluation, A-Precision, R-precision, and F-measures are used. The A-precision is the average precision when each relevant F-term is ranked to a test document. The R-precision indicates the precision when the top R relevant F-term is ranked in a test document, where R is the number of relevant categories. The F-measure is the average inverse of the combined recall and precision. The recall is the ratio of correct outputs to the total number of correct categories. The precision is the ratio of correct outputs to the total number of outputs.

Table 1 Number of themes and documents

	NTCIR-5	NTCIR-6	JPO
Theme	5	108	about 1,900
Test Document	2562	21606	Over 4.5 million

3. Basic Chi-square Statistic Weighting Method

To achieve a high-speed classification system, the method avoiding using a machine learning technique is proposed. The proposed method treats a word as a scalar value. In this section, the proposed patent classification method using the chi-square statistic weighting [6] is described in detail.

3.1 Preprocessing

A patent application composes five parts: bibliographical information (title of invention, application number, patent applicant, inventor, etc.), an abstract, a claim, a detailed description, and a brief description and schematic drawings. For this task, the proposed method uses only "abstract" and "claim". "ChaSen" is used for morphological analysis and only nouns are used in this method.

3.2 Chi-square statistic

Chi-square statistic weighting is a term weighting classification method, such as the TF-IDF [7] method. However, chi-square statistic weighting considers weights for words which do not appear in the documents as well as word which appear in them.

The chi-square statistic weighting applies the following equation to each word of each F-term:

$$\chi^2(x,y) = D(y,\alpha\gamma n) + D(x-y,\alpha(1-\gamma)n) + D(m-y,(1-\alpha)\gamma n) + D(n-x-(m-y),(1-\alpha)(1-\gamma)n) \quad (1)$$

where

$$\alpha = x/n,\ \gamma = m/n, \text{ and } D(o,e) = (o-e)^2/e \quad (2)$$

Each parameter can be estimated using the following 2x2 contingency table, which lists the F-terms of each word in the training patent documents.

Table 2 2x2 contingency table

	Word B	¬Word B	Subtotal
F-term A	y	x-y	x
¬F-term A	m-y	n-x-m+y	n-x
Subtotal	m	n-m	n

Here, y is the number of appearances of word B in the assigned F-term A in the document, x-y is the number of appearances in the document other than word B in the assigned F-term A, m-y is the number of appearances in the document of word B in other than the assigned F-term A, n-x-m+y is the number of appearances in the document of words other than word B in other than the assigned F-term A, x is the total number of the patent documents with assigned F-term A, m is the total number of appearances in the

document of word B, and n is the total number of the documents in each theme.

3.3 Classification

The ranking algorithm sums the calculated chi-square statistic weighting in each F-term in each word appearing in the test documents. After summing the chi-square statistic weighting, they are sorted in descending order:

$$R_j(D_j, F_i) = \sum_{W_j \in D_j} W_j(F_i) \quad (3)$$

where $W_j(F_i)$ is the chi-square statistic weighting of F-term i included in document j.

3.4 Evaluation at tentative test set

Table 3 shows the evaluation results for the average precision using the NTCIR-5 test collection as a tentative test set as well as the same results for the TF-IDF method as a baseline method of term weighting.

Table 3 Evaluation results of the chi-square statistic and baseline method at NTCIR-5 test collection

System	2B022	3G301	4B064	5H180	5J104
Chi-Square	0.4232	0.3171	0.4262	0.4837	0.3770
TF-IDF	0.3352	0.3053	0.4104	0.4648	0.3080

As show in Table 3, the chi-square statistic term weighting method performs well for all themes. In particular, the results of "2B022" and "5J104" are good. These themes are few training patent document.

The chi-square statistic term weighting classification method performs well even with few training data.

4. Improved Chi-square Method

The previous section indicates that the proposed method performs well compared with the baseline of the term weighting method. However, these evaluation results indicated that the proposed method had no advantage over the other methods used at NTCIR-5. Therefore, two improvements are applied to the proposed method.

In this section, the improvement of the chi-square statistic term weighting is described in detail.

4.1 N-gram chi-square statistic weighting

The N-gram is used to consider the co-occurrence relation between words. The N-gram chi-square statistic can be calculated in the same way as that for the words.

Table 4 N-gram 2x2 contingency table

	N-gram B	¬N-gram B	Subtotal
F-term A	y	x-y	x
¬F-term A	m-y	n-x-m+y	n-x
Subtotal	m	n-m	n

The value of document j is calculated as sums of the following two values:

· sums of chi-square statistic weights for words appeared in the document.

· sums of chi-square statistic weights for N-gram appeared in the document.

$$R_i(D_j, F_i) = \sum_{W_j \in D_j} W_j(F_i) + \alpha \sum_{N_j \in D_j} N_j(F_i) \quad (4)$$

where $N_j(F_i)$ is the N-gram chi-squared weighting of F-term i included in document j. In this paper, bi-gram (N=2) is used for the evaluation.

4.2 Weight emphasis for Words in F-term descriptions

Words in F-term descriptions are considered to be important words. Therefore, if the words in F-term descriptions appear in the test document, these word chi-square statistic weights are added to its weight:

$$R(D_j,F_i) = \sum_{W_j \in D_j} \begin{cases} \beta W_j(F_i) & \cdots\ W_j \notin L_i \\ W_j(F_i) & \cdots\ W_j \in L_i \end{cases} + \alpha \sum_{N_j \in D_j} N_j(F_i) \quad (5)$$

where L_i is word in F-term descriptions including F-term i.

4.3 Evaluation with the improved method

Table 5 shows the evaluation results for the average precision at NTCIR-5 test collection, as compared with the other methods considered herein.

In the Table 5, VSM, SVM and K-NN are method and result of other teams participated in classification subtask at NTCIR-5.

As show in Table 5, the improved method achieves accuracy as high as or better than the accuracy of other vector classification methods.

These results show that although term weighting and simple methods like the proposed method can obtain good results for a classified patent document.

Table 5 Evaluation results of the improved chi-square method at NTCIR-5 test collection

System	2B022	3G301	4B064	5H180	5J104
Proposed Method	0.4349	0.3956	0.4877	0.5451	0.4051
VSM	0.3142	0.1498	0.2196	0.2060	0.1516
SVM	0.3705	0.3568	0.4271	0.5080	0.3387
K-NN	0.4812	0.4091	0.5701	0.6222	0.4281

5. Evaluation of the NTCIR-6 test set

In this section, the evaluation results at NTCIR-6 of the proposed method are described and compared with all of the teams that participated in the classification subtask at NTCIR-6 patent retrieval task.

5.1 Evaluation results of the proposed system

The proposed method is applied to six runs with different parameters as shown in Table 6. The runs 1 through 3 use word chi-square statistic term weighting and bi-gram chi-square statistic term weighting. The runs 4 through 6 use word chi-square statistic term weighting, bi-gram chi-square statistic term weighting, and weights for words in the F-term description.

Table 6 System parameters

Run ID	N-gram α	F-term label β
NUT01	0.5	1
NUT02	0.8	1
NUT03	1.0	1
NUT04	0.5	2
NUT05	0.8	2
NUT06	1.0	2

Table 7 Evaluation results of NTCIR-6

Run ID	A-Precision	R-Precision	F-measure
NUT01	0.4039	0.3644	0.2391
NUT02	0.4053	0.3656	0.2366
NUT03	0.4065	0.3656	0.2357
NUT04	0.4090	0.3687	0.2469
NUT05	0.4101	0.3700	0.2432
NUT06	0.4100	0.3698	0.2413

The weight for each word in the F-term description is varied from 1 to 2, and the N-gram chi-square statistic weighting is varied from 0.5 to 1.0.

Table 7 shows the parameters and results for each system. For the proposed system, the best score was obtained by NUT5 using the N-gram chi-square statistic weight of 0.8 and the F-term label key word weight of 2.

5.2 Evaluation result of NTCIR-6

Figures 1, 2, and 3 show the results for all of the teams for A-Precision, R-Precision and F-measures, respectively. Six teams of 46 systems, including the team of the present study, participated in the workshop. The proposed method ranked 5th among all teams. However, there is no significant difference in A-Precision and R-Precision between the top four teams.

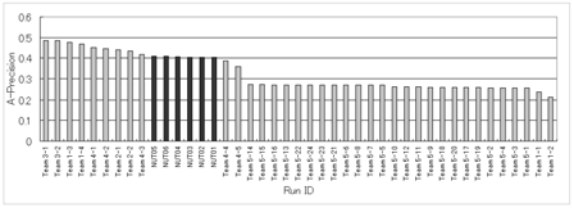

Fig. 1 Evaluation Result of A-Precision

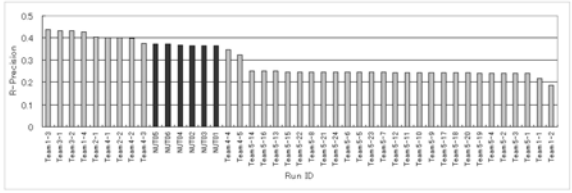

Fig. 2 Evaluation Result of R-Precision

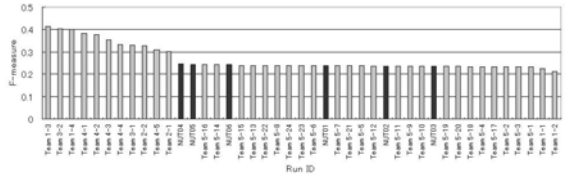

Fig. 3 Evaluation Result of F-measure

5.3 Discussion

Table 8 shows the classification methods of all teams. Most teams use a machine learning system or the vector classification method. These systems are assumed to require a long time to learn and classify all of the documents. The proposed system achieves fast classification.

In this task, the classification speed is not evaluated. The proposed method is evaluated in comparison with the vector classification method. During the preprocessing phase, the proposed method performs approximately 3.5 times faster than the vector classification method. In addition, during the classification phase, the proposed method performs approximately five times faster than the vector classification method.

Table 8 Classification method of the NTCIR-6 patent workshop

System	Method
Team 1	HSMV, SVM
Team 2	SVM, NB
Team 3	NB, Maximum Entropy
Team 4	K-NN
Team 5	K-NN
Proposed system	Chi-Squared

6. Conclusion

In the present paper, a high-accuracy and high-speed patent classification method is proposed for the F-term classification subtask.

The proposed method is a fast weighting method using the chi-square statistic term. The proposed method was applied to six systems in the classification subtask at NTICR-6 and the results of this application were evaluated. The results of accuracy evaluation were good, even though the best score was not obtained by the proposed method, confirming the effectiveness of the proposed method.

References

[1] Japan Patent Office. Administration of Patent 2006 Annual report.

[2] Makoto Iwayama, Atushi Fujii, Noriko Kando: Overview of Classification Subtask at NTCIR-5 Patent Retrieval Task, Proc. NTCIR-5 Workshop Meeting (2005).

[3] Y. Yang and X. Liu. A re-examination of text categorization methods. Proc. 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (1999).

[4] N. Cristianini and J. Shawe-Taylor. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press (2000).

[5] Makoto Iwayama, Atsushi Fujii, Noriko Kando, “Overview of Classification Subtask at NTCIR-6 Patent Retrieval Task”, Proceedings of the 6th NTCIR Workshop, 2007.

[6] Shinichi Morishita, Jun Sese: Traversing Itemset Lattices with Statistical Metric Pruning, Proc. ACM SIGACT-SIGMOD-SIGART Symp. On Database Systems (PODS), pp226-236 (2000).

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

[8] Shannon, C., “A Mathematical Theory of Communication,” Bell Syst. Tech. J.27.