

A Fast Patent Classification Method Using the Chi-square Statistic and Its Improvement

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

Kotaro Hashimoto
Nagaoka University of
Technology
1603-1 Kamitomioka-cho,
Nagaoka-shi, Niigata, Japan

Koji Mizumoto
Nagaoka University of
Technology
1603-1 Kamitomioka-cho,
Nagaoka-shi, Niigata, Japan

1. INTRODUCTION

The authors have been participating the Patent Retrieval Task at the NTCIR (NII Test Collection for IR Systems) Workshop [2]. We are focusing on the classification subtask, which categorizes target patent applications based on the F-term classification system.

In the previous NTCIR workshops, machine learning methods and vector classification methods provided good results, however, they are expected to require a long processing time when classifying a large number of patent documents. At the NTCIR-6, the authors proposed a high-speed classification method [1]. This paper introduces the method briefly and proposes its improvements. Evaluation results are also demonstrated.

2. BASIC CHI-SQUARE STATISTIC WEIGHTING METHOD

Our method treats a word as a scalar value. The weight for the word is determined based on chi-square statistic. Chi-square statistic weighting [3] is similar to the $TF \times IDF$ method. However, it considers weights for words which do not appear in the documents as well as word which appear in them. In addition to the simple chi-square statistic weighting, bi-gram is used to consider the co-occurrence relation between words. The bi-gram chi-square statistic can be calculated in the same way as that for the words.

The scoring algorithm basically sums the calculated chi-square statistic weighting in each F-term in each word appearing in the target documents.

Our system ranked 5th among six systems which participated the NTCIR-6 workshop.

3. IMPROVEMENTS OF THE METHOD

Two improvement methods are proposed. First improvement is focused on a fact that the tendency to the words is greatly different in each domain. It is suggested that high frequency words and one character words in the patent document adversely affect the classification accuracy. Those words are named “domain specific stop words.” The proposed method removes the domain specific stop words from the target document before computing scores.

Second improvement focuses on the “viewpoint” which is second-level of hierarchy of the category structure. Since

the basic method classifies a patent into F-term directly, the number of F-term candidates can be several hundreds and this leads deterioration in accuracy. A patent document is associated with multiple F-terms, however, these F-terms are subsidiary to two or three viewpoints. The proposed method firstly classifies the patent document into viewpoints, then classifies it into F-terms. Reducing the number of categories expects improving accuracy.

4. RESULTS AND CONCLUSION

The accuracy evaluation was done using the NTCIR-6 test set. Mean A-Precision (MAP) value for the proposed method was 0.4287. For comparison, the best MAP value for the basic chi-square method was 0.4101. The improved method achieves 4.5% better in MAP compared with the basic method.

For the speed comparison, a system based on the Vector Space Model (VSM) is implemented. The basic chi-square method performs approximately 3.5 times faster than the VSM-based method. The improved method requires additional computational efforts, however, it provides still faster processing speed.

From those results, it has been shown that the proposed methods achieved fairly good classification accuracy without losing the advantage of processing speed.

5. ACKNOWLEDGMENTS

The authors would like to thank the organizers of the NTCIR Patent Retrieval Task for their support with the patent test collections.

6. REFERENCES

- [1] K. Hashimoto and T. Yukawa. Term weighting classification system using the chi-square statistic for the classification subtask at ntcir-6 patent retrieval task. In Proceedings of NTCIR-6 Workshop Meeting, pages 385–389, 2007.
- [2] M. Iwayama, A. Fujii, N. Kando, and Y. Murakawa. Evaluating patent retrieval in the third ntcir workshop. Information Processing and Management, 42(1):207–221, 2006.
- [3] S. Morishita and J. Sese. Traversing itemset lattices with statistical metric pruning. In Proceedings of ACM SIGACT-SIGMOD-SIGART Symposium On Database Systems (PODS), pages 226–236, 2000.