# Chapter 32
# Application of Web Search Results for Document Classification

**So-Young Park, Juno Chang and Taesuk Kihl**

**Abstract** In this chapter, we propose a method applying Web search results to the document classification for the purpose of enriching the amount of the training corpus. For the query that will be submitted to a Web search engine, the proposed method generates the Web query based on the matching score between words in documents and the category. Experimental results show that the Web query based on the higher ranked words can improve the document classification performance while the Web query based on the lower ranked words makes worse the document classification performance.

**Keywords** Document classification · Web search results · Query generation

## 32.1 Introduction

Recently, a huge number of documents are produced and stored in digital archives [1]. Therefore, the document classification plays a very important role in many information management and retrieval tasks. It refers to the task of assigning a document to one pre-defined category. As described in the following Eq. (32.1), the document classification selects the category $c_i$, taking the highest matching probability for the given the unlabeled document $D$ where the category $c_i$ is one element of the category set $C$. In order to be easy to mathematically deal with the

S.-Y. Park · J. Chang · T. Kihl (✉)
SangMyung University, 20, Hongjimun 2-gil, Jongno-gu, Seoul, Korea
e-mail: tsroad@smu.ac.kr

S.-Y. Park
e-mail: ssoya@smu.ac.kr

J. Chang
e-mail: jchang@smu.ac.kr

H.-K. Jung et al. (eds.), *Future Information Communication Technology and Applications*, Lecture Notes in Electrical Engineering 235, DOI: 10.1007/978-94-007-6516-0_32,

document classification problem, most document classification methods represent the unlabeled document $D$ as the document vector $\vec{d}$ consisting of word occurrences in the document. Most elements of the document vector take zero value.

$$\arg\max_{c_i \in C} \; P(c_i \mid D) \;=\; \arg\max_{c_i \in C} \; P(c_i \mid \vec{d}) \tag{32.1}$$

Most digital documents are frequently updated, and the writers disseminate information and present their ideas on various topics [2]. Unlike news articles written by the well educated journalists, most digital documents such as weblogs tend to contain colloquial sentences and a slang language which misleads the classifier [3, 4]. Considering this cumbersomeness, some approaches have been proposed: the Naive Bayes-based approaches [4], the SVM based approaches [5], the knowledge-based approaches [6–8], the Maximum-Entropy based approaches [9–11], and Web based approaches [12].

In this paper, we propose the method applying the Web search results to the document classification. The rest of this paper is organized as follows. Section 32.2 will present the overview of the proposed method, and Sect. 32.3 will show some experimental results. Finally, the characteristics of the proposed method will conclude the paper in Sect. 32.4.

## 32.2 Application of Web Search Results for Document Classification

For the document classification, the proposed method assigns the relevant category to each document, as shown in Fig. 32.1. Given a document, for example, the proposed method assigns the category label such as A, B, C, or D to each document. Considering the data sparseness problem, the proposed method enriches the amount of the training corpus by cooperating with Web search engine.

Given the documents such as $d_1$ to $d_m$ with the corresponding categories such as $c_1$ to $c_l$ in the training set, the proposed method represents the documents as the document vectors consisting of the number of word occurrences in each document,

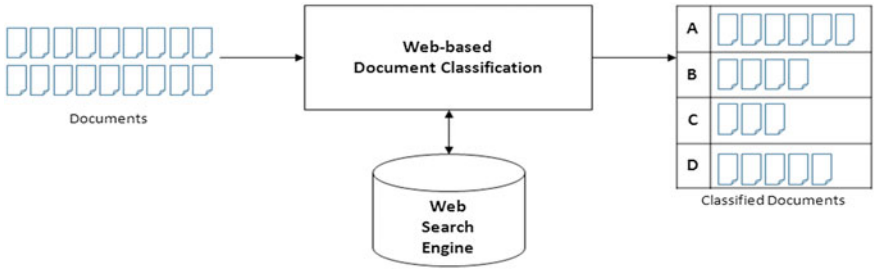


**Fig. 32.1** Document classification method using web search results

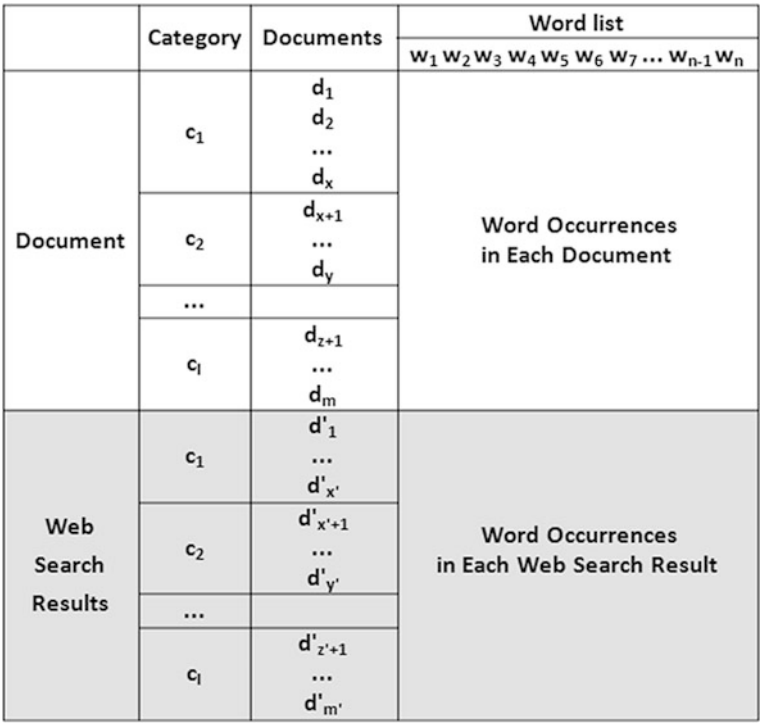


**Fig. 32.2** Training corpus adding the given documents to the web search results

as described in Fig. 32.2. According to the assumption that the larger training corpus is less troubled with the data sparseness problem, the proposed method adds the Web search results to the previously prepared training corpus. For the Web query generation, the proposed method first selects some useful words per category by using the Chi square statistics [13, 14]. The Chi square statistics of word feature $f_j$ in the category $c_i$ is defined as:

$$\chi^2(f_j, c_i) \approx \frac{N(AD - CB)^2}{(A + C)(B + D)(A + B)(C + D)} \tag{32.2}$$

where $A$ is the number of documents containing the word feature $f_j$ in the category $c_i$, $B$ is the number of documents containing the word feature $f_j$ in other categories rather than $c_i$, $C$ is the number of documents not containing the word feature $f_j$ in the category $c_i$, $D$ is the number of documents not containing the word feature $f_j$ in other categories rather than $c_i$, and N is the total number of documents. Each word feature $f_j$ was computed for every category, and the top $n$ word features with the higher Chi square statistics are used for the query candidates.

Finally, the proposed method sends the query to the Open API (Application Programmer Interface) of the Web search engine, and receives the snippet results

retrieved from the Web search engine. The proposed method assumes the snippet results of each query as one document.

## 32.3 Experiments

In order to prove the validity of utilizing the Web search results, we have tested the MALLET document classification package [9] with a mobile application description document corpus [15], which is divided into 90 % for the training set and 10 % for the test set. On the other hand, the proposed method is evaluated based on the evaluation criteria: precision, recall, and F-measure. Precision indicates the ratio of correct candidate categories from candidate categories predicted by the proposed document classification method. Recall indicates the ratio of correct candidate categories from the categories of the total documents in the corpus. F-measure indicates the harmonic mean of the precision and the recall. Because the document classification method predicts all categories of the given documents, the precision is the same as the recall. The baseline performance without any document classification method indicates 21 % because the documents corresponding to the *utility* capture roughly 21 % of the corpus.

**Table 32.1** Performance variation by the addition of web search results

| | Mallet | | | Mallet + Web search results | | |
|---|---|---|---|---|---|---|
| | Precision | Recall | F-measure | Precision | Recall | F-measure |
| 0 | 41.43 | 41.43 | 41.43 | 41.43 | 41.43 | 41.43 |
| 10 | 41.43 | 41.43 | 41.43 | 41.48 | 41.48 | 41.48 |
| 20 | 41.43 | 41.43 | 41.43 | 41.65 | 41.65 | 41.65 |
| 30 | 41.43 | 41.43 | 41.43 | 41.74 | 41.74 | 41.74 |
| 40 | 41.43 | 41.43 | 41.43 | 41.71 | 41.71 | 41.71 |
| 50 | 41.43 | 41.43 | 41.43 | 41.11 | 41.11 | 41.11 |
| 60 | 41.43 | 41.43 | 41.43 | 41.66 | 41.66 | 41.66 |
| 70 | 41.43 | 41.43 | 41.43 | 41.40 | 41.40 | 41.40 |
| 80 | 41.43 | 41.43 | 41.43 | 41.77 | 41.77 | 41.77 |
| 90 | 41.43 | 41.43 | 41.43 | 41.60 | 41.60 | 41.60 |
| 100 | 41.43 | 41.43 | 41.43 | 41.74 | 41.74 | 41.74 |
| 110 | 41.43 | 41.43 | 41.43 | 41.66 | 41.66 | 41.66 |
| 120 | 41.43 | 41.43 | 41.43 | 41.74 | 41.74 | 41.74 |
| 130 | 41.43 | 41.43 | 41.43 | 41.88 | 41.88 | 41.88 |
| 140 | 41.43 | 41.43 | 41.43 | 41.80 | 41.80 | 41.80 |
| 150 | 41.43 | 41.43 | 41.43 | 41.51 | 41.51 | 41.51 |
| 160 | 41.43 | 41.43 | 41.43 | 41.12 | 41.12 | 41.12 |
| 170 | 41.43 | 41.43 | 41.43 | 41.09 | 41.09 | 41.09 |
| 180 | 41.43 | 41.43 | 41.43 | 41.46 | 41.46 | 41.46 |
| 190 | 41.43 | 41.43 | 41.43 | 41.20 | 41.20 | 41.20 |
| 200 | 41.43 | 41.43 | 41.43 | 41.34 | 41.34 | 41.34 |

The *mallet* performance indicates the recall of the baseline document classification method *mallet*, learned from the given training set without any Web search results. The *mallet + Web search results* performance indicates the recall of the document classification method *mallet*, learned from the training set adding an equal number of Web search results for each category. Table 32.1 describes that the *mallet + Web search results* method is a little bit more effective than the *mallet* method. Also, Table 32.1 shows that the performance does not always increase according to the addition of the Web search results; because the characteristics of mobile application description documents is too different from the characteristics of Web search results, which is too sensitive to the Web search queries. Besides, the performance does not generally increase too much by adding the Web search results; since most category prediction results in the test set are biased towards few categories corresponding to many documents while the Web search results help the categories with few documents.

## 32.4 Conclusion

In this paper, we propose a method applying Web search results to the document classification. The proposed method has the following characteristics. The proposed method enriches the amount of the training corpus by applying Web search results. Also, the proposed method cooperates with Web search engine by generating the Web query that will be submitted to a Web search engine. Experimental results show that the Web query based on the higher ranked words can improve the document classification performance while the Web query based on the lower ranked words makes worse the document classification performance. For future works, we want to propose a document classification method considering balancing the number of documents per category, and compare the effects of two methods adding Web search results. Unlike the already proposed method always add the same number of the Web search results to each category; the hopefully proposed method does not add any Web search results for the categories having many documents, while the method adds many Web search results for the categories with a few documents.

**Acknowledgments** This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2012R1A1A3013405).

## References

1. Nyberg K, Raiko T, Tinanen T, Hyvnen E (2010) Document classification utilising ontologies and relations between documents. In: Proceedings of the 8th workshop on mining and learning with graphs, Washington, pp 86–93

2. Ayyasamy RK, Tahayna B, Alhashmi S, Eu-gene S, Egerton S (2010) Mining wikipedia knowledge to improve document indexing and classification. In: 10th international conference on information science, signal processing and their applications, pp 806–809
3. Ferreiraa R, Freitasa F, Britob P, Meloa J, Limaa R, Costab E (2013) RetriBlog: an architecture-centered framework for developing blog crawlers. Expert Syst Appl 40(4):1177–1195
4. Park S, Kim CW, An DU (2009) E-mail classification and category re-organization using dynamic category hierarchy and PCA. J Inf Commun Convergence Eng 7(3):351–355
5. Baharudin B, Lee LH, Khan K (2010) A review of machine learning algorithms for text-documents classification. J Adv Inf Technol 1(1):4–20
6. Rubin TN, Chambers A, Smyth P, Steyvers M (2012) Statistical topic models for multi-label document classification. Mach Learn 88(1–2):157–208
7. Lu G, Huang P, He L, Cu C, Li X (2010) A new semantic similarity measuring method based on web search engines. WSEAS Trans Comput 9(1):1–10
8. Jialei Z, Hwang CG, Jung GD, Choi YK (2011) A design of K-XMDR search system using topic maps. J Inf Commun Convergence Eng 9(3):287–294
9. McCallum AK (2002) MALLET: a machine learning for language toolkit. http://mallet.cs.umass.edu
10. Berger A, Pietra SD, Pietra VD (1996) A maximum entropy approach to natural language processing. Comput Linguist 22(1):39–71
11. Lim JH, Hwang YS, Park SY, Rim HC (2004) Semantic role labeling using maximum entropy model. In: Shared task of the fourteenth conference on computational natural language learning
12. Samarawickrama S, Jayaratne L (2011) Automatic text classification and focused crawling. In: Sixth international conference on digital information management (ICDIM), pp 143–148
13. Yang Y, Pedersen JO (1997) A comparative study on feature selection in text categorization. In: 14th international conference on machine learning, pp 412–420
14. Seki K, Mostafa J (2005) An application of text categorization methods to gene ontology annotation. In: 28th annual international ACM SIGIR conference on research and development in information retrieval, pp 138–145
15. Kihl T, Chang J, Park SY (2012) Application tag system based on experience and pleasure for hedonic searches. Convergence Hybrid Inf Technol Commun Comput Inf Sci 310:342–352