


# A Sample Extension Method Based on Wikipedia and Its Application in Text Classification

Wenhao Zhu<sup>1</sup> · Yiting Liu<sup>1</sup> · Guannan Hu<sup>1</sup> · Jianyue Ni<sup>1</sup> ·  
Zhiguo Lu<sup>2</sup> 

Published online: 8 February 2018

© Springer Science+Business Media, LLC, part of Springer Nature 2018

**Abstract** Text classification is a topic in natural language processing that is particularly useful for Internet information processing. Methods based on supervised learning require a large amount of manually annotated training samples. The annotation of training samples is time consuming, and performance relies heavily on the quality of the training samples. This paper presents a text classification method based on sample extension. The extension is based on the correlation of the labeled sample data and the concepts in Wikipedia. Combined with the rich link relationships between concepts, we selected appropriate articles from Wikipedia to expand the training sample set. By introducing the large amount of rich semantic concept pages that are contained in Wikipedia along with links that are related to different pages, our approach enhances the performance and generalization of the classifier. Experiments demonstrate that the performance of the method proposed in this paper is better than that of both supervised and semi-supervised methods.

**Keywords** Text classification · Semi-supervised learning · Sample extension · Wikipedia

## 1 Introduction

With the rapid development of the Internet, the big data era is coming. Text data are increasing at an explosive rate; thus, it has become difficult to develop an information retrieval application with high efficiency. The sorting of text data is also essential, and one of the effective solutions is classification. As a key technology for natural language processing (NLP), text classification has been continuously developed and has widespread

---

✉ Zhiguo Lu  
Luzg@staff.shu.edu.cn

<sup>1</sup> School of Computer Engineering and Science, Shanghai University, Shanghai, China

<sup>2</sup> Library of Shanghai University, Shanghai, China

implications. The common methods used in text classification include naive Bayes [19], support vector machine [15, 20], K-nearest neighbor [10], and neural networks [21].

Most of these classification methods are examples of supervised learning. To enhance the performance of the classifier, it is necessary to improve the quantity and quality of samples. In fact, it is easy to obtain unlabeled data. However, annotating samples is time consuming and expensive, which is likely to lead to bottlenecks. To solve this problem, semi-supervised learning, which is a comprehensive learning method consisting of a combination of labeled and unlabeled samples, was developed. With the guidance of the labeled samples and the effective information of the unlabeled samples, this method can improve the performance of the classifier.

One of the problems with supervised learning is that it does not take full advantage of the extensive knowledge in unlabeled data; thus, the text classification is limited to the contents of the original labeled data. This problem causes all samples to be limited in a particular domain. In addition, semi-supervised learning requires a large amount of unlabeled data, but the improper selection of unlabeled data may affect the classification performance.

The use of data from a rich corpus with a reasonable screening mechanism to expand the training sample set is more likely to solve the problem of text classification. Wikipedia, which is an artificial calibration source, contains a wealth of articles, and its links between the articles constitute a large knowledge network. Wikipedia can assist text classification to overcome the difficulty of classifying unlabeled samples, reduce the burden and cost of manually annotating samples, and achieve the goal of improving the generalization and practicality of text classification. The use of Wikipedia for inductive transfer for text classification is effective, as is using Wikipedia to automatically construct a thesaurus of concepts to enhance previous approaches to text classification [1, 18].

This paper presents a text classification method based on Wikipedia sample extension. The correlation between text and concepts has become an important part of data mining for Wikipedia. By calculating the correlation between concepts and the labeled sample, as well as the rich links between concepts, our approach selects appropriate articles from Wikipedia to build an extended sample set. On this basis, the generalization of the classifier is improved. The experimental results show that the text classification performance improved after sample expansion.

The remainder of this paper is organized as follows. Section 2 introduces the theory and work associated with establishing the model. Section 3 introduces the structure of the model. Section 4 describes the experimental process and the analysis of the results. Section 5 summarizes the article.

## 2 Related Work

With the development of information technology, to effectively manage and utilize the large amount of data from the Internet, information retrieval and mining based on content have become areas of concern. Text classification is an important data analysis method in information retrieval, information mining and other areas of research. Text classification automatically determines the categories of text. The main goal of this study is to improve classification performance.

Automatic text classification technology was proposed as early as the 1950s. Luhn proposed the use of word frequency statistics, which are mainly used for automatic

classification in this field, in a pioneering study. Since then, a large number of researchers have studied related fields of text classification [7]. Kspark, Salton, Jones and other scholars in this field performed significant work in this area [16]. Since the 1990s, automatic text classification technology based on machine learning and statistics has gradually become the mainstream of text classification and is commonly used in supervised text classification [2]. In traditional text classification, the text is labeled by one or more given category tags according to the feature terms contained in the text based on a pre-defined text classification model system. A general text classification process includes text pre-processing, feature extraction, text representation, classifier training and testing processes. At this stage, the research of text classification focuses on two aspects: improving the classifier model and improving the sample set. Regarding the improvement of the sample set, the traditional supervised algorithm requires a significant artificial annotation corpus, which results in a large workload, is time consuming and is associated with certain technical demands. To solve the problems caused by the lack of training samples, text classification based on semi-supervised learning was developed. Semi-supervised learning uses a large amount of unlabeled data combined with a small amount of labeled data to train a model. Shahshahani and Landgrebe began researching semi-supervised learning in 1994. Since then, semi-supervised learning has become a growing research area [5]. Chapelle et al. proposed nuclear-based semi-supervised learning [4]. Wajeed proposed a semi-supervised text categorization method based on K-nearest neighbor (KNN) that uses different similarity measurements and different vector generation techniques to improve classification accuracy [17]. Li et al. proposed a self-trained SVM algorithm [12]. Pavlinek proposed a semi-supervised latent Dirichlet allocation (LDA) text classification method that used the semi-supervised LDA theme model to identify a topic [14]. This paper focuses on improving the sample set based on Wikipedia. A classification algorithm based on semi-supervised learning is equivalent to massive training sample expansion. The qualities of the sample sets are the key to improving the accuracy of classifiers. Therefore, the appropriate selection of unlabeled data, reasonable labeling and the use of a screening mechanism for building candidate training sample sets are crucial for extending training samples and improving the performance of the classifier.

### 3 Classification Method Based on Wikipedia Sample Extension

#### 3.1 Data Preprocessing

The original text must be expressed in a form with simple preprocessing. The initial data set is very different in terms of the format and text type of the Chinese corpus in this paper. To achieve better classification results, a series of pretreatments are required:

1. Simplification of traditional words. The tool used for the simplification of traditional words is the Open Chinese Convert open source project [3].
2. Word segmentation. The present study uses the word segmentation tool jieba, which is an open source tool [11].
3. Removal of punctuation, numbers, and stop words. The filter uses a stop word list containing 2112 Chinese words. The stop word list includes functional words (e.g., virtual words, modifiers, conjunctions, and adverbs) and some other words that contain the lowest lexical meaning.

In particular, it is also necessary to use regular expressions to filter out complex formats of the Wikipedia corpus in Chinese, i.e., to filter out the references, comments, external links and other unnecessary information. At this point, the preprocessing of complex text has been completed.

### 3.2 Wikipedia Sample Extension (WSE)

To expand the samples properly, the present study utilized the initial labeled dataset to annotate a large, semantically rich corpus and selected data from that corpus to extend the training set by an appropriate screening mechanism. Articles in Wikipedia are rich, and they describe separate concepts clearly. Therefore, it is suitable as a candidate set for sample extension. In particular, articles have a large number of links, which contain semantic concept relationships. Because the links are marked by artificial means, their semantics are less ambiguous. There are thousands of levels of link relationships, and each article has an average of 34 links. The size of the dataset is approximately 3.3 GB, and the number of the articles is 1188008 as of 2016.10. Thus, according to the size and structure of the background knowledge, Wikipedia is a very good dataset for sample extension based on correlation. This paper proposes a sample extension method that is based on Wikipedia and correlation calculation. On this basis, we extract and use the rich link information and articles of Wikipedia to expand samples. The WSE consists of four steps: step 1: calculate the correlation of every concept in Wikipedia for each category; step 2: screen data based on correlation and construct a sample extension candidate set; step 3: select the appropriate article corresponding to a concept to expand the training sample; and step 4: train a better classifier. The roadmap for the sample extension is shown as follows Fig. 1:

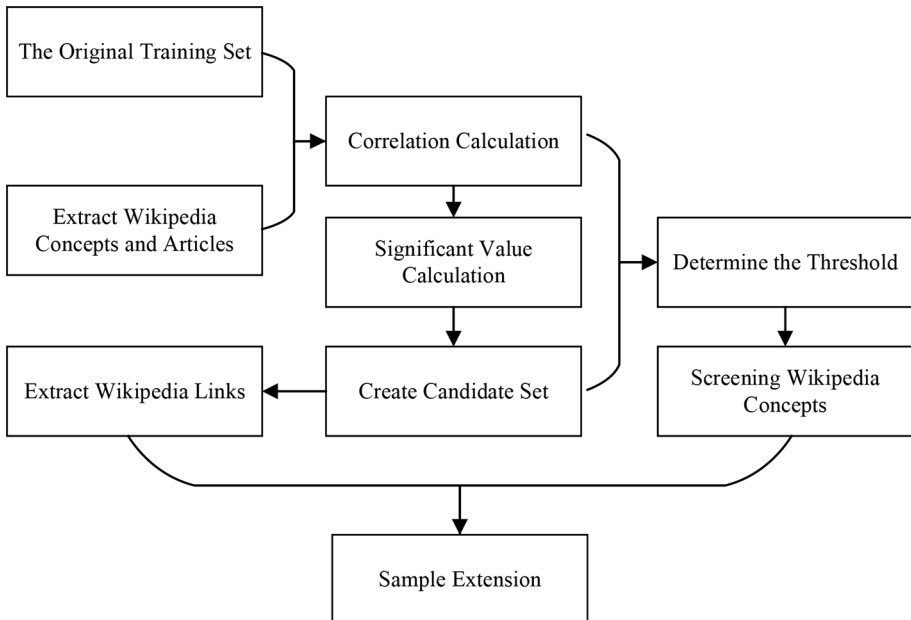


Fig. 1 Based on Wikipedia’s sample extension

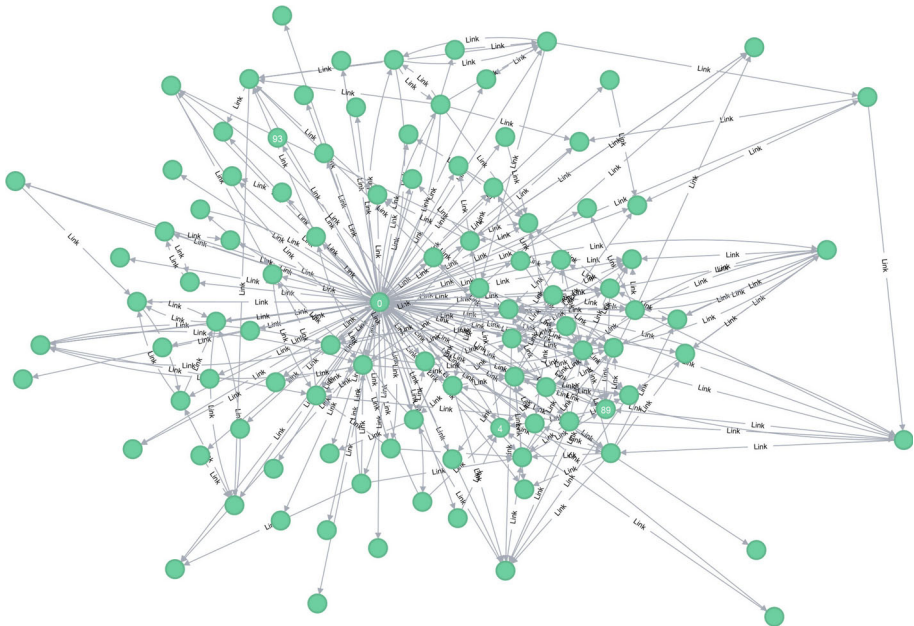
### 3.2.1 Construct Network Graph of Wikipedia

There are numerous links in Wikipedia's articles, and most of the links between concepts represent their close semantic relationships. We can build tight complex network graphs for concepts based on the rich links in Wikipedia. Concepts and their links are extracted from Wikipedia. As shown in Fig. 2, each concept is a link node, and each link is a relationship between nodes. Thus, a linked graph of concepts is constructed, and the linked graph is directed. For example, the article of "economics" has a link to the concept "stock", and the node "economics" has a directed side to the node "stock". In particular, there are many special pages called ambiguous pages in Wikipedia. These pages should be eliminated and marked as ambiguous nodes.

Before starting the sample extension, the nodes that point to each other with the ambiguous nodes should be screened out according to the Wikipedia network graph. The main reason for this step is that the nodes that point to ambiguous nodes are ambiguous concepts; their corresponding articles in Wikipedia cover different categories. Such contents are not suitable for extending samples for text classification. This step reduces the amount of data to be processed and improves the efficiency of subsequent processing. Finally, the Wikipedia network graph contains approximately one million nodes and approximately 40 million edges.

### 3.2.2 Correlation Calculation

Measuring concepts of semantic correlations is the basic goal in the field of NLP [8]. In the field of text classification, the semantic correlations of different terms can greatly improve its effect [9]. Extensive research has been conducted on the calculation of text similarity,



**Fig. 2** Part of Wikipedia network graph

which has been applied in many fields [13]. This paper uses TF-IDF as a basis for calculating the correlations of Wikipedia concepts and sample categories. TF-IDF is a simple, intuitive, fast method for processing text feature weightings, and it has been widely used in text processing [16]. Specific practices and formulas are illustrated as follows:

$$\begin{aligned} tfidf &= TF * IDF \\ &= (count_{w,D} / count_{all,D}) * \log(size_{all} / size_w) \end{aligned} \quad (1)$$

$$(TAvg_A) = \left( \sum_{i=1}^S tfidf_i \right) / S \quad (2)$$

Formula (1) calculates the TF-IDF for each Wikipedia concept and each document. In this formula,  $count_{w,D}$  represents the number of times that word  $w$  appears in document  $D$ , and  $count_{all,D}$  represents the sum of the occurrences of all words in document  $D$ .  $size_{all}$  represents the total number of documents, and  $size_w$  represents the number of documents that contain word  $w$ . TF indicates the frequency at which concept  $w$  appears in document  $D$ , which reflects the importance of a concept relative to a document. IDF is actually the reciprocal of the DF (document frequency), and DF is the total number of documents in which the word appears in the entire corpus. Formula (2) calculates the average TF-IDF for each category, where  $S$  is the total number of documents for category  $A$ , and  $TAvg_A$  is the average TF-IDF for those documents.

### 3.2.3 Build a Sample Extension Candidate Set

According to the correlations of Wikipedia concepts and each category, it is necessary to screen out concepts that represent a category to significantly expand samples. The concepts that are not highly relevant to this category but that are highly relevant to other categories should be excluded. The specific method is to use formula (3) to calculate the significant value of each Wikipedia concept for each category:  $TValue_I$  ( $I = A, B$ )

$$(TValue_A) = \frac{TAvg_A}{\prod TAvg_J + \varepsilon} (J = B, C, \dots) \quad (3)$$

In formula (3),  $\varepsilon$  is set to 0.00001 to avoid a denominator of 0.  $TValue_A$  represents the importance of the concept in category  $A$  compared with other categories. When the value of  $TAvg_A$  is larger and the  $TAvg$  of other categories is smaller, the value of  $TValue_A$  is greater, i.e., the significance of the concept for category  $A$  is also greater. Therefore, the concept is the better representation of category  $A$ . For each concept, the  $TValue$  of all categories should be ranked, the most significant category should be selected to mark it, and it should be placed into the candidate set of that category. In formula (4), if  $I = A$ , then the word is classified as belonging to category  $A$ , and its  $TValue_{max}$  is recorded.

$$(TValue_{max}) = \text{Max } TValue_I (I = A, B, C, \dots) \quad (4)$$

Finally, for each category, concepts should be sorted by their  $TValue_{max}$ , concepts should be selected in order, and the articles corresponding to those concepts should be placed into the training dataset to complete the expansion of the training dataset. The ratio of sample expansion is based on a large number of comparative experiments in this paper.

### 3.2.4 Wikipedia Sample Extensions with Links (WSE-L)

Wikipedia concept links constitute a tight complex network structure, and they represent close relationships between concepts. From the analysis of the rich edges in the network graph of Wikipedia in Sect. 3.2.1 of our paper, it can be observed that there is a high correlation between the conceptual nodes that point to each other. For example, the concept node of “sports” and the concept node of “martial arts” point to each other. If they point to each other, then we mine new concepts that have interlinked relationships in the graph with the concept that is in the original candidate sample set and add their articles to the extended sample set. Therefore, based on the WSE, this paper proposes an enhanced sample extension method, Wikipedia sample extensions with links (WSE-L), i.e., the use of rich Wikipedia links to extract closely related articles to expand the sample further.

### 3.2.5 Wikipedia Sample Extensions with Limited Links (WSE-LL)

Although the links between Wikipedia concepts symbolize the correlation between them, enhancement of the sample extension based on the links results in a significant increase in the number of training sample sets. For this purpose, we present Wikipedia sample extensions with limited links (WSE-LL). WSE-LL adds a limited condition to the method WSE-L; the new concept must have appeared in the sample extension candidate set obtained by WSE. For example, the concept “economics” is linked to “stock”, and “stock” appears in the candidate set. Then, we extract the article of “stock” into the extended sample set. Finally, under limited conditions, we can achieve a more accurate sample expansion by using rich Wikipedia links.

## 4 Experiment and Analysis

### 4.1 Experimental Data

The training samples in this paper consist of labeled and unlabeled samples. Among them, the labeled samples are from Sina news datasets, and the unlabeled samples are from Wikipedia. Web crawler is used to obtain the Sina online news for the categories of technology, sports, economics, military, and entertainment. To place this information into categories and improve the training effect of the model, 300 news items are selected for each category as the unlabeled sample set. Wikipedia contains millions of concept pages, and each page contains rich links. All of the concepts and their corresponding articles are extracted from Wikipedia, constituting the unlabeled sample set, and a graph (node, relation) is constructed according to the links between those concepts. The test dataset is Sohu news and the text classification corpus of Fudan University, including articles in the categories of technology, sports, economics, military, and entertainment, and the number of items in each category is 300.

## 4.2 Experimental Process

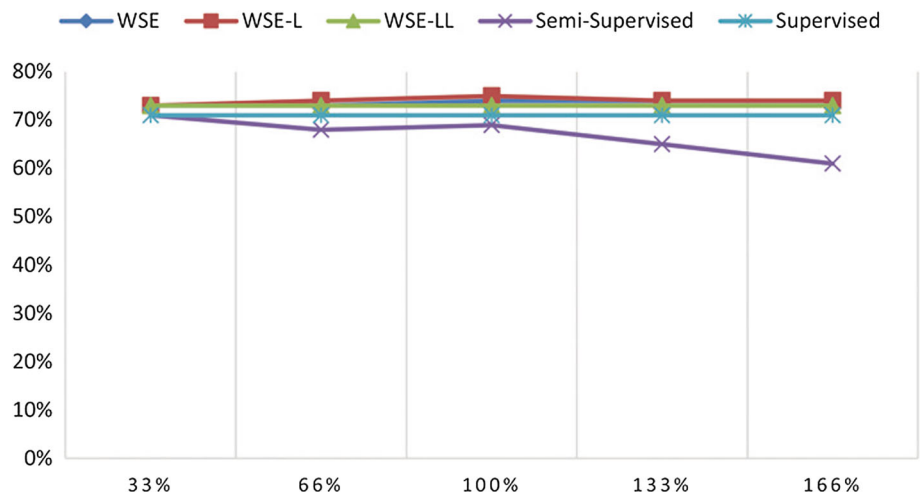
### 4.2.1 Sample Expansion

First, the text dataset is obtained by preprocessing the news datasets. Second, according to Sect. 2, a network graph of Wikipedia is constructed, and the concept of ambiguity based on the graph is removed. Based on the significant value of the Wikipedia concept calculated by the sample extension algorithm in Sect. 2, the corresponding concept is marked and added to the candidate set. The corresponding articles in each category are pre-processed and added to the training set. In addition, WSE-L, which is based on the WSE, extracts the articles of interrelated concepts from the graph and adds these marked articles to the training sets. With the condition of the sample extension candidate sets, WSE-LL filters articles with conceptual links, marks them and adds them to the training sets.

### 4.2.2 Training Classifier

The text classification models are constructed with the classification algorithm. In this paper, the classifiers are naive Bayes, support vector machine and random forest. By training these models, a large number of documents are expressed as a number of subject information. This method uses the full training samples to analyze the related parameters. The classifier models in this paper are implemented in Python, which uses the tool Sklearn.

Before training the classifier, the feature selection in this paper uses the characteristic frequency to select the characteristic words. The first 1500 words are selected for each category, a unique id is defined for each word, and a feature dictionary is built. According to the feature dictionary, the training matrix is calculated. Each row represents a training text in the matrix, and the head of each row is the total number of different words in the classifier dictionary. The matrix is generated in the form “id: count”, and the count is the frequency of the word in the text. Each datum is separated by spaces. The number of columns in the matrix is equal to the sum of the words in the dictionary. The number of rows in the matrix is equal to the sum of articles in the entire training set.



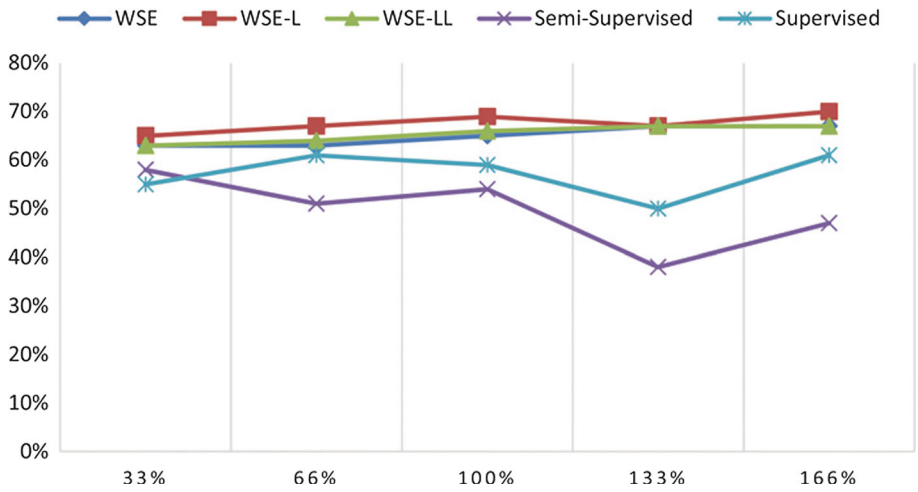
**Fig. 3** The relationship of the accuracy of Sohu news and the proportion of sample expansion



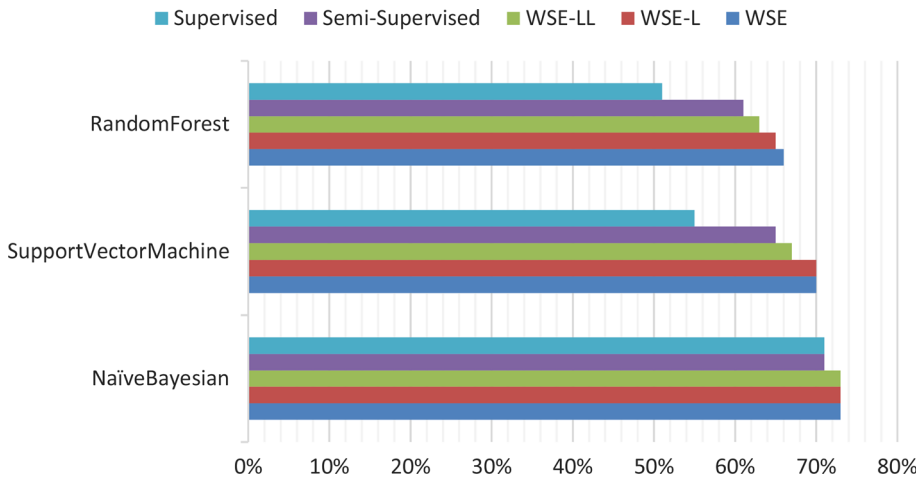
### 4.3 Evaluation Criteria

To evaluate the performance of the text classification, three general evaluation criteria are used: accuracy, F-measure and ROC. The standard formulas are shown in formula (5) and formula (6):

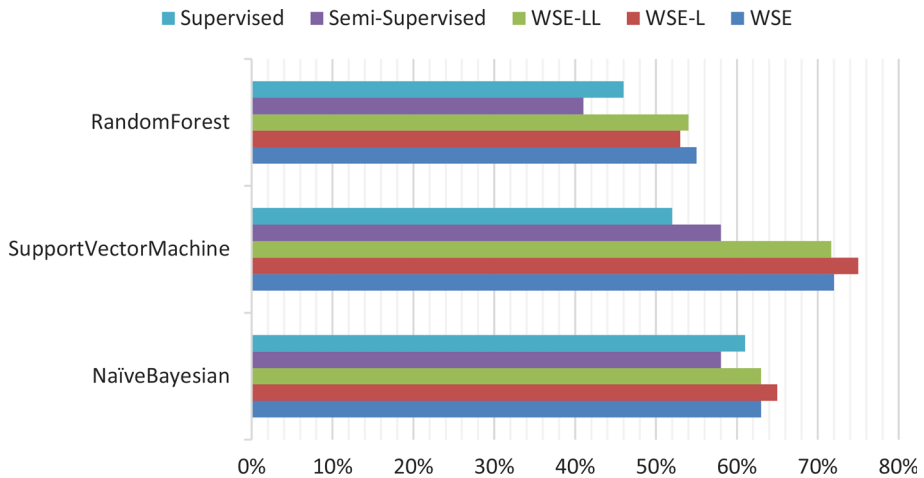
$$Accuracy = \frac{\text{Number of news being classified correctly}}{\text{Total number of news}} \tag{5}$$



**Fig. 4** The relationship of the accuracy of text from Fudan University and the proportion of sample expansion



**Fig. 5** The accuracy of Sohu news



**Fig. 6** The accuracy of text from Fudan University

$$F\text{-measure}_A = \frac{\text{Recall}_A * \text{Precision}_A * 2}{\text{Recall}_A + \text{Precision}_A} \quad (6)$$

Accuracy is the probability that all samples are correctly classified. Recall is the probability that the samples in this category are correctly classified, i.e., the ratio of the sum of documents that are classified correctly and the total number of test document sets for this category. Precision is the probability that the classifier is classified correctly, i.e., the ratio of the sum of documents that are correctly classified to the total number of documents in this category. F-measure is the harmonic average of recall and precision, reflecting the comprehensive indicators. The values of the evaluation criteria are between 0 and 1; if the value is closer to 1, accuracy, F-measure, recall and precision will be higher. The ROC score is the area under the plot created by the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.

#### 4.4 Results and Analysis

Our method was compared with semi-supervised text classification, which is based on topic information proposed by Dorado and Ratté [6]. Their method is to use the least amount of labeled data to accelerate the creation of the corpus for a specific classification process. Moreover, our method was also compared with text classification, which is based on Wikipedia's artificial annotation. The accuracy and F-measure for each dataset are calculated, and the performance of the classifier is analyzed.

To test the effect of the classifier for the sample extension, we apply the dataset as described in Sect. 4.1 to the methods of WSE, WSE-L, WSE-LL, supervised text classification and semi-supervised text classification. Additionally, we also increased the number of unlabeled training samples and observed the accuracy trend of those methods. The X-axis indicates the proportion of sample extensions. Using Sohu news as a test set, the experimental results are shown in Fig. 3). In addition, the corpus of Fudan University is also used as a test set, and accuracy comparisons are shown in Fig. 4. As shown in Figs. 3 and 4, the accuracies of WSE, WSE-L and the WSE-LL are higher than those of the

traditional text classification method. When the number of samples reaches a certain value, the accuracy increased nearly 10%. As shown in the figures, the performance of WSE-L is better than that of other methods for those datasets. The main reason for this result is that the expansion set obtained by WSE without limit contained more words; their articles contain the content of several categories. In contrast, WSE-L can play an excellent role in corpora that contain diverse texts. In addition, to verify the effectiveness of our method, we run our methods on different classifiers multiple times. These classifiers are random forest, support vector machine and naive Bayes. Based on the expansion of 100 samples, we observe the performance of each method on different classifiers. The test data also contain Sohu news and text from Fudan University. Figures 5 and 6 show the accuracy of the classification results. Figures 7 and 8 show the F-measure of the classification results. As

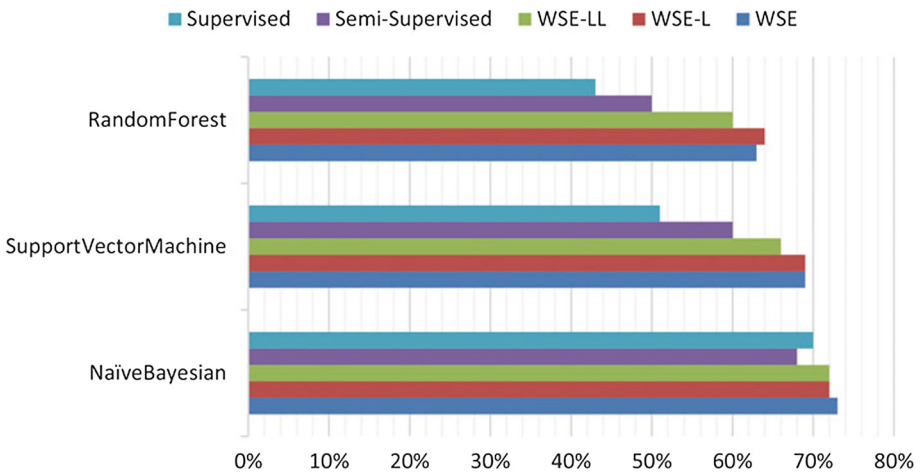


Fig. 7 The F-measure of Sohu news



Fig. 8 The F-measure of text from Fudan University

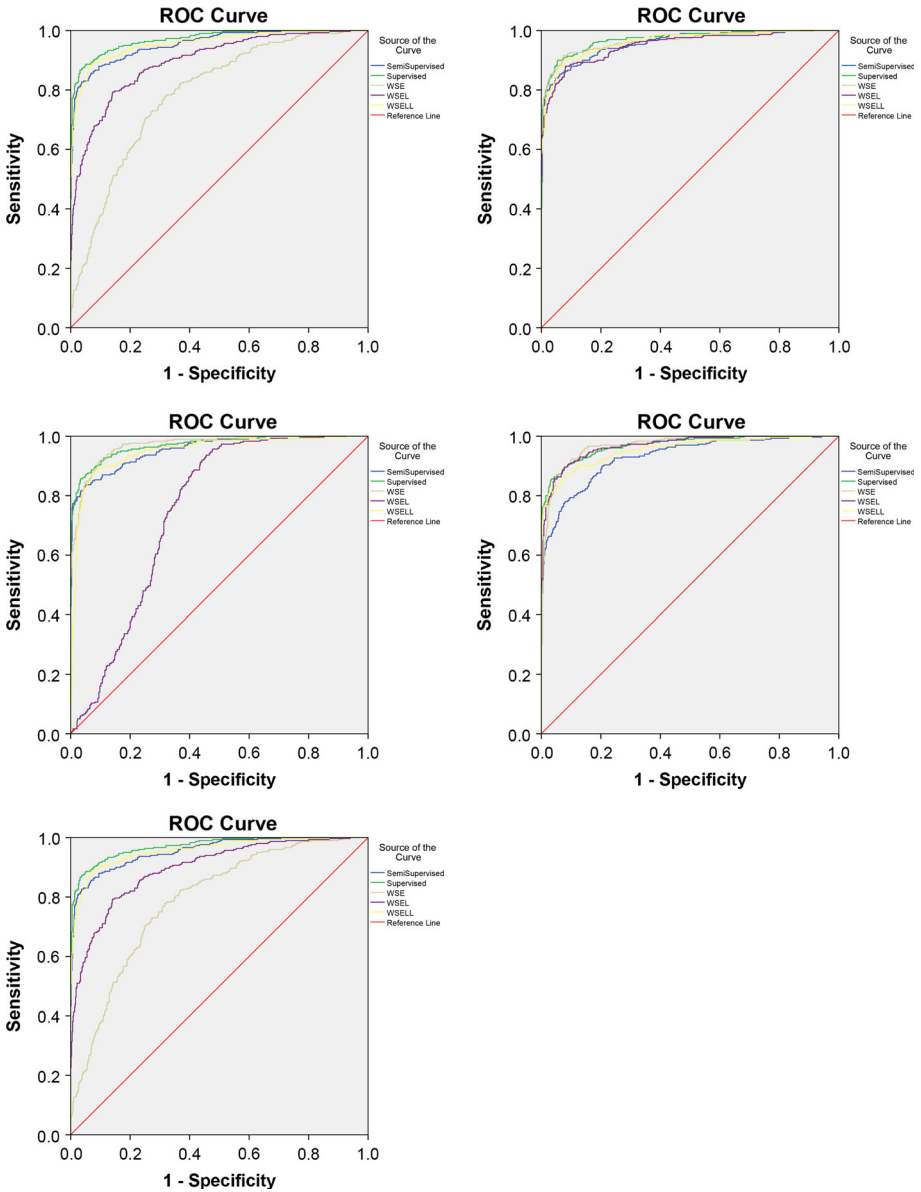
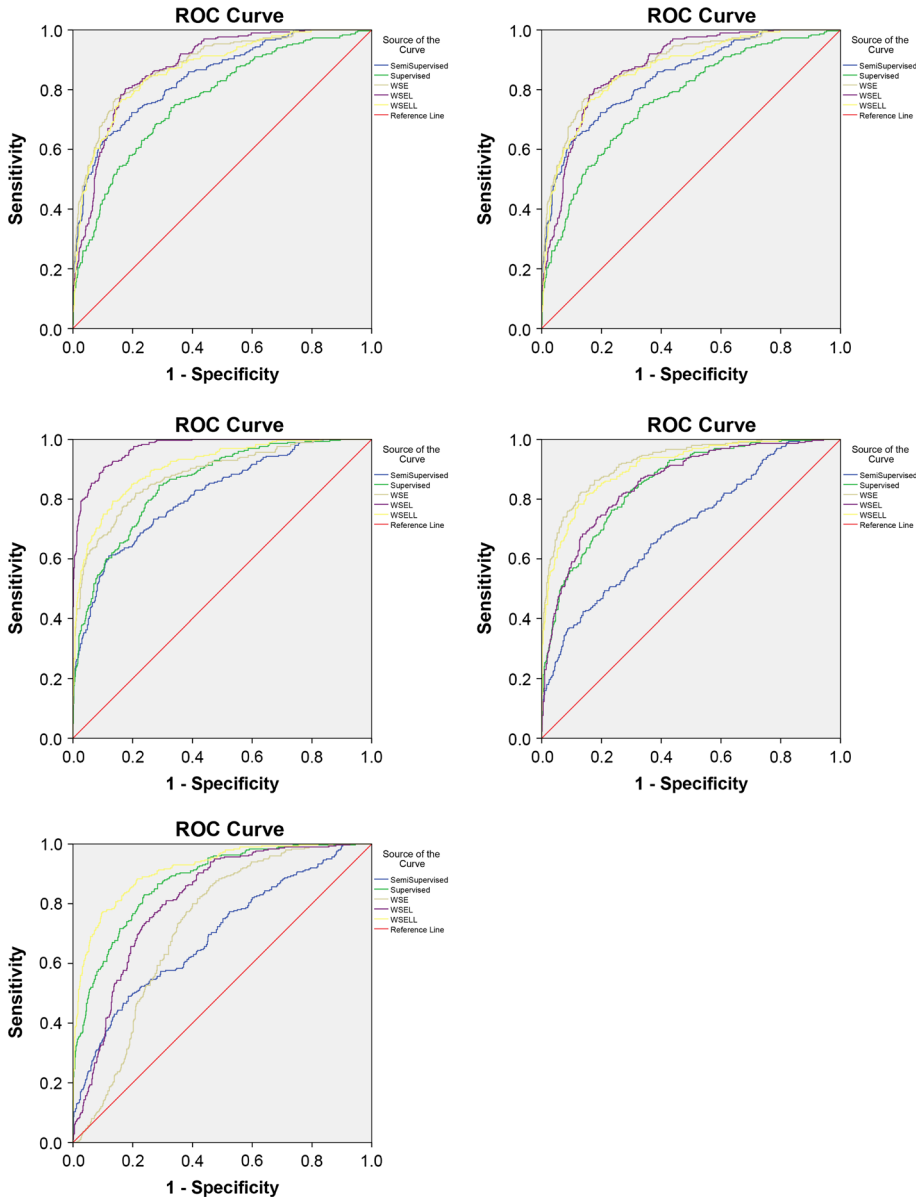


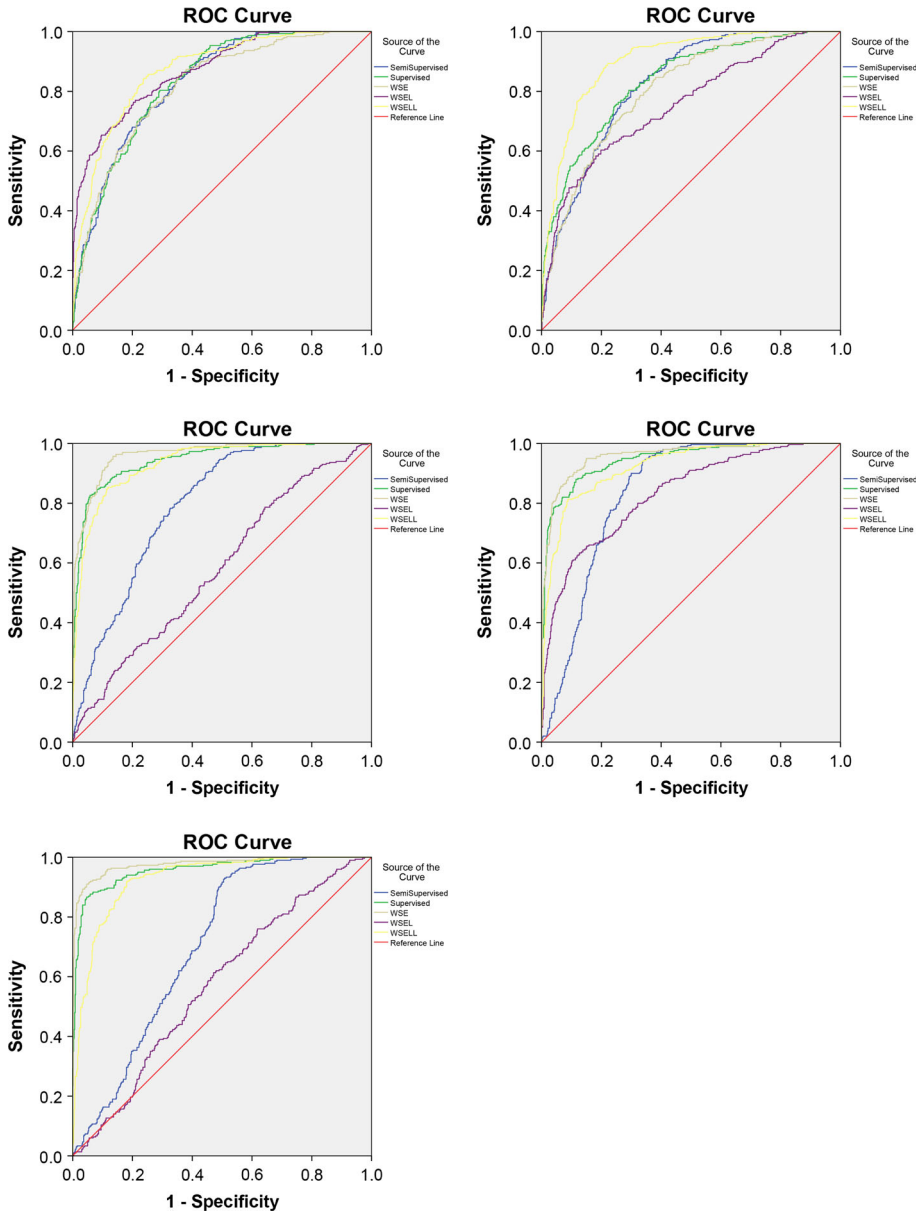
Fig. 9 The ROC of sports classification from Fudan University

shown in the figures, the performance of our method proposed in this paper is relatively stable and effective. Compared with the semi-supervised text classification, our method greatly improves the performance and generalization of text classification. The analysis of the characteristics of the two different test sets indicated that our method is more generalized. The approach performs well in the classification of news, as well as in the classification of journal articles and presentations. According to the analysis of the data characteristics and F-measure, the “economic” category of text is concise and timeliness is



**Fig. 10** The ROC of economic classification from Fudan University

strong. Our method makes the classifier more suitable for use in other scenes. However, for those categories of samples that are highly professional or samples with relatively rich semantics, the improvement of the classifier based on our method is not very clear. Thus, the method proposed in this paper is more suitable for training samples with limited content and uncertain applications. Finally, the test dataset of Fudan University and the tool of SPSS are used to create the ROC curves, as shown in Figs. 9, 10 and 11. These figures include the experiments that have different proportions of sample expansion. Using



**Fig. 11** The ROC of technology classification from Fudan University

our method, the ROC curve is closer to the upper left corner of the coordinates, which is more intuitive to better reflect the performance of our method.

## 5 Conclusion

This paper calculates the correlations used to identify Wikipedia concepts and filters the appropriate text to expand the training samples used for text classification. We proposed several approaches (WSE, WSE-L, and WSE-LL) to expand the samples. The results show that WSE-LL is superior to supervised learning approaches. The addition of Wikipedia expands the quantity and quality of training samples properly. In particular, we use its rich links to enhance the sample expansion effectively. The results also show that our method improves the performance of text classification, which is a very viable method. Regarding future work, we can consider the following directions. (1) Using parallelization techniques to accelerate classification speed. Due to the large amount of corpora, the improvement of computational efficiency is an important aspect of our future research. (2) Applying the Wikipedia sample extension to other fields of NLP (e.g., emotional analysis).

**Acknowledgements** The work of this paper is partially supported by the National Natural Science Foundation of China (Nos. 61572434, 61303097).

## References

1. Banerjee, S. (2007). Boosting inductive transfer for text classification using wikipedia. In *Sixth International Conference on Machine Learning and Applications, 2007 (ICMLA 2007)* (pp. 148–153).
2. Bijalwan, V., Kumar, V., Kumari, P., & Pascual, J. (2014). Knn based machine learning approach for text and document mining. *International Journal of Database Theory and Application*, 7(1), 61–70.
3. BYVoid: Openc (2014). <https://github.com/BYVoid/OpenCC>. Accessed 10 Nov 2016.
4. Chapelle, O., & Zien, A. (2005). Semi-supervised classification by low density separation. In *AISTATS* (pp. 57–64).
5. Dópido, I., Li, J., Marpu, P. R., Plaza, A., Dias, J. M. B., & Benediktsson, J. A. (2013). Semisupervised self-learning for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 51(7), 4032–4044.
6. Dorado, R., & Ratté, S. (2016). Semisupervised text classification using unsupervised topic information. In *FLAIRS*.
7. Galán-García, P., De La Puerta, J. G., Gómez, C. L., Santos, I., & Bringas, P. G. (2015). Supervised machine learning for the detection of troll profiles in twitter social network: Application to a real case of cyberbullying. *Logic Journal of IGPL*, 24(1), 42–53.
8. Harispe, S., Ranwez, S., Janaqi, S., & Montmain, J. (2013). Semantic measures for the comparison of units of language, concepts or instances from text and knowledge base analysis. arXiv preprint [arXiv: 1310.1285](https://arxiv.org/abs/1310.1285).
9. Harispe, S., Sánchez, D., Ranwez, S., Janaqi, S., & Montmain, J. (2014). A framework for unifying ontology-based semantic similarity measures: A study in the biomedical domain. *Journal of Biomedical Informatics*, 48, 38–53.
10. Jiang, S., Pang, G., Wu, M., & Kuang, L. (2012). An improved k-nearest-neighbor algorithm for text categorization. *Expert Systems with Applications*, 39(1), 1503–1509.
11. Junyi, S. (2017). <https://github.com/fxsjy/jieba>. Accessed 25 Nov 2016.
12. Li, Y., Guan, C., Li, H., & Chin, Z. (2008). A self-training semi-supervised SVM algorithm and its application in an EEG-based brain computer interface speller system. *Pattern Recognition Letters*, 29(9), 1285–1294.
13. Low, Y., & Zheng, A. X. (2012). Fast top-k similarity queries via matrix compression. In *Proceedings of the 21st ACM international conference on information and knowledge management* (pp. 2070–2074).
14. Pavlinek, M., & Podgorelec, V. (2017). Text classification method based on self-training and lda topic models. *Expert Systems with Applications*, 80, 83–93.
15. Ramírez, J., Górriz, J., Salas-Gonzalez, D., Romero, A., López, M., Álvarez, I., et al. (2013). Computer-aided diagnosis of alzheimers type dementia combining support vector machines and discriminant set of features. *Information Sciences*, 237, 59–72.

16. Van Dongen, B., Dijkman, R., & Mendling, J. (2013). Measuring similarity between business process models. In *Seminal contributions to information systems engineering* (pp. 405–419). Berlin: Springer.
17. Wajeed, M.A., Adilakshmi, T. (2011). Semi-supervised text classification using enhanced KNN algorithm. In *2011 World Congress on information and communication technologies (WICT)* (pp. 138–142).
18. Wang, P., Hu, J., Zeng, H. J., & Chen, Z. (2009). Using wikipedia knowledge to improve text classification. *Knowledge and Information Systems*, 19(3), 265–281.
19. Wang, X. Z., He, Y. L., & Wang, D. D. (2014). Non-naive bayesian classifiers for classification problems with continuous attributes. *IEEE Transactions on Cybernetics*, 44(1), 21–39.
20. Yoshikawa, Y., Iwata, T., & Sawada, H. (2014). Latent support measure machines for bag-of-words data classification. In *Advances in neural information processing systems* (pp. 1961–1969).
21. Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. In *Advances in neural information processing systems* (pp. 649–657).



**Wenhao Zhu** received his Ph.D. in computer science from Zhejiang University, China in 2009. He is currently an associate professor in the School of Computer Engineering and Science in Shanghai University. His main research topics include information extraction, information retrieval, and digital Library.



**Yiting Liu** received her B.S. degree from Anhui University of Finance and Economics in 2013. She is now studying towards a master's degree in the School of Computer Engineering and Science of Shanghai University. Her main research topics are information extraction and information retrieval.





**Guannan Hu** received his B.S. in software engineering from Donghua University in 2009, and M.S. degree in computer science from Shanghai University, and now is a candidate Ph.D. in School of Computer Engineering and Science in Shanghai University. His main research topics include reinforcement learning and deep learning.



**Jianyue Ni** received his M.Sc. degree in advanced software engineering from the University of Sheffield. He is now a Ph.D. candidate in the School of Computer Engineering and Science of Shanghai University. His main research topics include machine learning, ontology and high-throughput material calculation.



**Zhiguo Lu** is now studying for Ph.D. at School of Computer Engineering and Science of Shanghai University and working in Shanghai University library as a deputy director. His main research topics are digital library, high performance computing and high performance information service.