IJAMML 1:1 (2014) 35-54 October 2014
ISSN: 2394-2258
Available at http://scientificadvances.co.in

A NEW METHOD FOR CLASSIFYING CHINESE TEXT BASED ON SEMANTIC TOPICS AND DENSITY PEAKS

Yewang Chen* **and Jixiang Du**

College of Computer Science and Technology, Huaqiao University, Xiamen, P. R. China

Abstract

This paper presents a new classification method for Chinese texts based on semantics topics and density peaks. The main motivation of this work is that the most existing text classification methods fail to deal with Chinese Web texts, because of the sparsity and irregularity of these Web texts. The novel method proposed, comes up with an idea of gaining the real semantics of text as features; these semantics should be stable and abstract that express the high hierarchical semantics behind the text, and can be used to differentiate different text categories. Therefore, firstly, BaiduBaike is used to extract the semantic topics as the real semantics from a text. Secondly, a clustering method is applied for finding the density peaks of each text category. Finally, the text is classified by the distances among the text and density peaks. This method deal with Chinese Web short texts well with fewer training data. The conducted experiments have shown that our method is promising, especially in the case of training data is not enough or processing Chinese Web texts.

Keywords: semantic topic, BaiduBaike, density peaks, Chinese phrase.

*Corresponding author.

E-mail address: ywchen@hqu.edu.cn (Yewang Chen).

2010 Mathematics Subject Classification: 68.
Submitted by Jianqiang Gao.
Received August 11, 2014

1. Introduction

With the advent of information age, Web information manifests an explosive growth throughout internet. There are billions of the Web texts emerge everyday by variety of ways, such as e-commerce, online communication, forum and chat messages etc. However, it is difficult to retrieve the useful data for users, for most of these texts have nothing to do with them. Therefore, process and make use of these resources effectively becomes increasingly important in many Web applications. During the past decades, learning to classify Web texts and documents using machine learning methods has been intensively studied, and it attracts more and more people, thank to it provides proper useful data for different users. Many machine learning methods [1], [2] are proposed, such as Naive Bayes, maximum entropy, KNN, and SVM; most of these methods have been applied in many cases [1] and achieved satisfactory results.

However, Chinese text is difficult to process for two reasons: The first is that the basic unit of Chinese is not hanzi, but Chinese phrase, and the second one is that there is no natural delimiter to depart Chinese phrases. Things are even worse in the case of processing the Chinese Web texts, for there are many short length and irregular expressions, as well as novel catchwords in these Web texts. They do not provide enough word co-occurrence or shared context for a good similarity measure. It brings great challenges for normal machine learning methods to recognize and process these expression, and most of existing machine learning methods fail to achieve desire accuracy. In our previous works [3], we have the opinion that the key reasons they fail are the following. The first reason is that hanzis and Chinese phrases are only one of the expressions of real semantics, i.e., given semantic topic, there exists a great number of Chinese phrases to express the semantic, it is impossible to enumerate all of them in a finite training data; the second one is that new Chinese phrases appear in the Web everyday, and it is difficult for us to update the training data in time.

Inspired by the idea mentioned above, we have the point of view that process Chinese texts by real semantics is more reliable than by superficial and variable expressions, i.e., hanzis or Chinese phrases. Therefore, we propose a new method to classify Chinese text based on the features of the real semantics of text, instead of hanzis or Chinese phrases directly. Specifically, this method utilizes BaiduBaike to extract the semantic topics of a text, and applies a clustering method [17] to detect density peaks for different text categories, and then distances between the text and density peaks are used to classify the Chinese text. Our experiments have demonstrated that this method is promising.

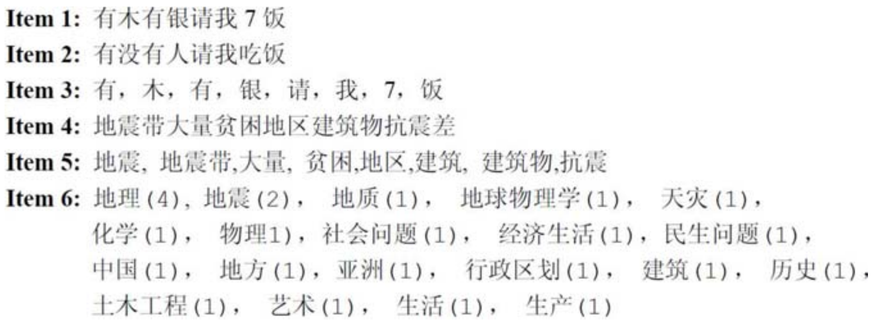
The rest of this paper is organized as follows. In Section 2, the related works is introduced. The principle of our method is presented in Section 3. The classification processes is presented in Section 4. Section 5 shows our experiments and analysis. The final section is our conclusion.

2. Related Works

In the past decades, there is a great deal of studies have been carried out in the field of automatic text classification (categorization), which is one of the research hotspots in the area of text mining and information retrieval. The main methods for text classification discussed in Sebastiani [1] fall within the machine learning paradigm, for example, SVM and KNN. Sebastiani also discussed the matters of document representation, classifier construction and the evaluation in detail. Su et al. [2] introduced a survey on the state-of-the-art of text categorization and highlighted the challenging issues and research trends.

Topic modelling has been a popular machine learning method for text mining in recent years, represented by pLSA [4] and LDA [5]. The idea of topic modelling is to create a probabilistic generative model for the text documents in the corpus. It is suitable for uncovering the latent structure behind the document collections, and provides new ways for searching, browsing, and summarizing large archives of texts. Xu and Wang [6] presented the development of topic models from LSI to pLSI and LDA in detail, and focused on the internal relationship among them.

Most of these methods mentioned above have been applied widely and achieved satisfactory results. In the case of processing Chinese, however, it is more complicated for these machine learning methods to understand and process for some reasons introduced in the first section, i.e., there is no natural delimiter in Chinese texts, and the basic unit for Chinese to express semantics is not hanzi, but Chinese phrases. Therefore, the first step to deal with Chinese text is Chinese word segmentation, which is also an active research area in the Chinese language processing community. Many technologies have been proposed, word-based method played the dominant role in the early work, and recently the character-based tagging method [7], [8] appeared in Bakeoff-2005 and became the most popular method for its remarkable effect. ICTCLAS [9] is a famous framework for Chinese word segmentation. Based on HMM model, it achieves high accuracy rate in processing regular Chinese texts. However, it still fails to achieve desired effect in the case of processing Chinese Web short texts, because of the sparsity and irregularity of these texts. For example, the Item 1 in Figure 1 is irregular and it means 'is there anybody want to invited me to have dinner? in Chinese as shown in Item 2. The word segmentation result by ICTCLAS is shown in Item 3. Obviously, it is total wrong for ICTCLAS can not understand these irregular expressions. In order to deal with this problem, Xia et al. [10] proposed a method to convert these irregular and short texts to standard language by phonetic mapping model.



Item 1: 有木有银请我 7 饭

Item 2: 有没有人请我吃饭

Item 3: 有，木，有，银，请，我，7，饭

Item 4: 地震带大量贫困地区建筑物抗震差

Item 5: 地震, 地震带,大量, 贫困,地区,建筑, 建筑物,抗震

Item 6: 地理(4), 地震(2)， 地质(1)， 地球物理学(1)， 天灾(1)，
化学(1)， 物理1)，社会问题(1)， 经济生活(1)，民生问题(1)，
中国(1)， 地方(1)，亚洲(1)， 行政区划(1)， 建筑(1)， 历史(1)，
土木工程(1)， 艺术(1)， 生活(1)， 生产(1)

Figure 1. Some Chinese text examples.

There also have been extensive studies and rapid progresses in Chinese text categorization for recent years. Song et al. [11] introduced a new model for semantic representation of Chinese texts by gaining extra information from wikipedia, in order to retain the contextual information for each word with a large extent. An improved labelled-LDA model was proposed by Jiang et al. [12]. In this method, the labels have two components, one is local topic and the other is shared topic, and they were treated as key features for classification. Li et al. [13] put forward another labelled-LDA method to improve the traditional LDA. Integrating the class information, this method introduced a new algorithm to determine the number of latent topics for each class. Teng [14] introduced a method based on CRFs for Web short text, character-based and character-tagging were used to extract features. Li [15] provided a tool for classifying Chinese texts based on SVM and KNN, the ICTCLAS was used for Chinese word segmentation. All these methods introduced in this paragraph work well on the regular texts. However, there are two disadvantages, the first is large scale of training data is needed in the training phase; the second one is in the case of texts are sparse, short and irregular, these methods can not achieve the desire accuracy.

3. The Principle of our Method

3.1. The disadvantages of existing methods

The first step to process Chinese texts is word segmentation in word-based methods, and then these methods train and classify texts by different mathematical model. Clearly, the result of Chinese words segmentation will severely affect the final result. For example, 'Apple IV' is totally different from 'four bags of apples', in spite of both of them contain 'Apple'. Take 'NBA' and 'National Basketball Association' for another example, they are the same, while the similarity between them is zero. In order to deal with these difficulties, new ways should be proposed to process these problem.

We put foreword an opinion in our previous works [3] that the Chinese words or phrases are only the manifestation of the real semantic. For a given semantic topic, there are great number of Chinese phrases or words that are semantic relevance to it. Therefore, it is impossible to list all Chinese phrases or words in a limited training data. It means that Chinese phrases or words are alternative for a semantic, in another words, they can be replaced by different Chinese phrases with the same meaning. While, the semantic topic and the topic relation of these endless Chinese phrases or words are relatively stable. For example, 'Apple', 'Banana', 'Tomato'... are all a kind of 'Fruit'. 'Fruit' is more stable than 'Apple', 'Banana', and 'Tomato', for it is a abstract concept in the area of agriculture, i.e., this concept indicates the deep meaning that expresses the high hierarchical relationship in this area. From this point of view, it is feasible for machine learning methods to process Chinese information by training few and well chosen training data, provided, we could extract abstract semantics behind the Chinese texts with stable relationship among these semantics.

BaiduBaike is an open and free knowledge base in Chinese Web, it provides comprehensive, accurate, and complex information about each Chinese phrase. Furthermore, it keeps up with the hot spots and network catchwords. There are some advantages of BaiduBaike as the following:

(1) Comprehensiveness: There are about 3.4 million Baike-phrases in BaiduBaike. Overall, it covers all domains of the society, even Web catchwords, such as ‘OMG’.

(2) Real-time: BaiduBaike is woven into the events of the day, it creates a Baike-phrase in time when a hotspot event happens, such as ‘MH370’, and it also updates the Baike-phrase with the progress of the event.

(3) Relationship: There are rich relationship among Baike-phrases, so that it is easy for a Baike-pharse to find other related Baike-phrases.

(4) Variety: There are some varieties or synonymies for a Baike-phrase, such as ‘C.R.’ (Cristiano Ronaldo, a famous soccer player) is the same as ‘Cristiano Ronaldo’.

The current Chinese segmentation tools, such as ICTCLAS [9] can not deal with catchwords, which makes side-effects on understanding and classifying Chinese Web text. Therefore, we have the point of view that BaiduBaike is suitable to be an infrastructure for Chinese text mining by providing real time, accurate, and rich information.

3.2. Baike-phrase, semantic topic and Chinese text

In BaiduBaike, Baike-phrase is the basic unit. Each Baike-phrase has signature, reference, open class, main body, related phrases, and extension. These different parts show information about the Baike-phrase from different aspects, and it is noticed that the ‘open class’ shows the abstract knowledge or relationship about a Baike-pharse, that is, it expresses the deep semantic of the Baike-pharse. Therefore, we, intuitively, have the opinion that the ‘open class’ can be used to express the real semantics for Chinese texts.

Definition 1. Semantic topic [3]: An open class of a Baike-phrase is a semantic topic.

For example, the Baike-phrase 'Tank' has the following semantic topics, namely, military, land army, weapon, armored car. Obviously, these semantic topics are the deep meaning, i.e., the semantic, behind 'Tank'. We have proposed some basic points of view in our previous works [3] as the following:

(1) Baike-phrases are only the manifestation of the real semantic: For a given semantic topic, there are many, sometimes are infinite, Chinese phrases that are related to it, the training data can not list all of them.

(2) Semantic topics are connotation: They are stable, abstract knowledge that express the high hierarchical semantics behind the text.

(3) Statistical regularity: The more important of a semantic topic in a Chinese text, the more Baike-phrases related to it there are; two Chinese texts with similar semantics have similar semantic topics.

3.3. The basic idea

Inspired by the points of view mentioned above, we come up with the idea that in order to classify Chinese texts, it is necessary to find the real semantics behind them. Therefore, intuitively, we can use BaiduBaike to map a Chinese text into a set of semantic topics, then take these semantic topics as features and set a proper mathematic model for classifying Chinese texts. There are some advantages of this way, the first is that we do not deal with the variable features, i.e., Chinese phrases or words, directly; the second is we do not need great mounts of training data any more; it is easy process Chinese Web texts.

Algorithm 1: Detecting Candidate-Phrases

```
Input 1: Baike-phrases prefix MAP: $map$
Input 2: A string $T$
Output: All candidate Baike-phrases
1. $n$ = length of $T$
2. result = null
3. For ($i$ = 1 to $n$)
4.     For ($j = i + 1$ to $n$)
5.         if $T[i, j]$ is not a key in $map$
6.             return to Step 3
7.         else
8.             if $T[i, j]$ is a Baike-phrase
9.                 result = result $\cup T[i, j]$
10.            end if
11.        end if
12.    end for
13. end for
14. return result
```

Therefore, in this paper, we propose a method for classifying Chinese texts based on BaiduBaike and clustering density peaks. This method requires fewer training data than any other existing methods, and our experiments have shown that it performs well on Chinese Web texts and regular Chinese texts. The main processes are the following:

(1) Detect all possible Baike-phrases in a Chinese text.

(2) Extract all semantic topics of the text.

(3) Find the density peaks for each category.

(4) Classify a text by the distances between the text and those peaks.

4. The Classification Processes

4.1. Detect Baike-phrases in a Chinese text

Definition 2. Candidate-phrase [3]: Given T is a text, $T_{i,j}$ is a string that starts from the i-th char to the j-th char of T, $T_{i,j}$ is a candidate-phrase if $T_{i,j}$ was a Baike-phrase.

In our previous works, we have build a prefix base [16] for all Baike-phrases, and provided an algorithm to detect all candidate-phrases effectively based on this prefix base, as shown in Algorithm 1.

4.2. Extract semantic topics

Definition 3. Semantic relevance [3]: Given e is a semantic topic, w is a Baike-phrase, and T is a Chinese text. We say w is semantic relevance to e if e was one of the semantic topics of w, and say T is semantic relevance to e if there was at least one candidate-phrase of T that is semantic relevance to e.

And now, given a Chinese text, it is easy for us to extract all relevant apparent semantic topics of it. Take Item 4 in Figure 1, for example, the candidate-phrases are shown in Item 5, and then semantic topics are extracted as the Item 6 shows. The number behind a topic in the set is the total number of candidate-phrases that are semantic relevance to the topic in the text. For example, there are 4 candidate-phrases are semantic relevance to the first topic. Obviously, these semantic topics in the set reveal the deep meaning behind the original text. In the following steps of our method, we use these semantic topics instead of Chinese phrases as features of the original text for classification.

Given a space of text $(T_1, T_2, \dots, T_N)$, and supposed there are M numbers of different semantic topics in the whole space, all these semantic topics consist of a topics dictionary $DIC = (topic_1, topic_2, \dots, topic_M)$. Then for an arbitrary text sample T, we transfer it into a semantic topic vector as follows.

Definition 4. Semantic topic vector: For a Chinese text *T*, semantic topic vector is defined as $TopicVec(T) = (v_1, v_2, \ldots, v_M)$, where v_i is the statistical times of $topic_i$ that appears in *T*.

4.3. Finding density peaks

In this part, a new clustering method [17] in machine learning field is used to find density peaks in one category, then these peaks are used as key samples that represent the category. For an unlabelled Chinese text *T* and a category *c*, we extract features of *T*, then calculate the distance between *T* and each peak of *c*, and the shortest one is chosen as the distance from *T* to the category *c*. The detail is shown as the following.

The clustering method used in this paper is based on the idea that cluster centers are characterized by a higher density than their neighbours and by a relatively large distance from points with higher densities [17]. It recognizes the clusters, spots and excludes outliers automatically regardless of their shape and of the dimensionality of the space. For each data point *i*, it computes its local density ρ_i and its distance d_i from points of other higher density. Both these quantities depend only on the distances $d_{i,j}$ between data points, which are assumed to satisfy the triangular inequality. The local density ρ_i of data point *i* is defined as

$$\rho_i = \sum_j \chi(d_{i,j} - d_c), \tag{1}$$

where d_c is a cutoff distance, and $\begin{cases} \chi(x) = 1, & \text{if } x < 0, \\ \chi(x) = 0, & \text{else,} \end{cases}$

δ_i is measured by computing the minimum distance between the point *i* and any other point with higher density:

$$\delta_i = \begin{cases} \min\limits_{j:\rho_j > \rho_i} (d_{i,j}), \text{ if } \exists\ j \text{ s.t. } \rho_j > \rho_i, \\ \max\limits_j (d_{i,j}), \qquad \text{otherwise.} \end{cases} \tag{2}$$

The points with high δ and ρ value are called as peaks that have higher densities than other points. Each peak point can be treated as a cluster center, and a point is assigned to the same cluster as its nearest neighbour peak. The peaks in category *c* are chosen as

$$PEAKS(c) = \{peak_1, peak_2, \ldots, peak_V\}, \tag{3}$$

where *V* is the number of peaks, and $\forall peak \in PEAKS(c)$, s.t. both ρ_{peak} and δ_{peak} are top 5% value.

4.4. Classify a text

Given a Chinese text *T*, we extract its semantic topics as features. The distance between T_i and T_j is defined as $d(T_i, T_j) = \|TopicVec(T_i) - TopicVec(T_j)\|$. Given a set of training text categories $C = \{c_1, c_2, \ldots\}$. For each category, we find peaks by the clustering method [17] introduced in the previous subsection, and then for an unknown Chinese text *T*, the distance between *T* and category c_k is defined as

$$dist(T, c_k) = \min_{peak \in PEAKS(c_k)} d(T, peak). \tag{4}$$

The distance between *T* and c_k indicates the probability of *T* belongs to c_k. The lower the distance is, the higher probability it belongs to category c_k. Therefore, *T* could be classified into the category that has the shortest distance to *T*

$$category(T) = \arg\min_{c \in C} dist(T, c). \tag{5}$$

5. Experiments and Analysis

In this section, a series of experiments are conducted in order to evaluate our proposed method, providing comparisons with the most popular existing methods, such as KNN, SVM, LDA, and CRF.

5.1. Experimental data

We collect 3,341,626 Baike-phrases, and build 9,959,704 prefixes of these Baike-phrases. The total number of semantic topics is 546,276. The following of this part presents our data set for experiments, and the detail is shown in Table 1.

Table 1. The description of training and test data

Training/test	Data set 1	Data set 2	Data set 3	
Traffic	2000/715	30/715	Women	500/9500
Phy-Edu	1000/482	30/482	Phy-Edu	500/9500
Military	2000/435	30/435	Military	500/9500
Medicine	1500/543	30/543	News	500/9500
Political	2500/701	30/701	Travel	500/9500
Education	1500/550	30/550	Education	500/9500

Data set 1 is a regular Chinese documents collection that downloads from Fudan NLP [18], we select 6 categories and 13,926 Chinese documents for experiment.

Data set 2 is a subset of Data set 1. We select the same 6 categories as Data set 1, but each category has only 30 documents for training data, and the number of documents for test is also the same as Data set 1. This data set is used for testing our method in the case of training data is not enough.

Data set 3 is a Chinese Web short texts set that comes from SogouC [19] (full version), and we only extract the title of each document as Web short text. We also select 6 categories, each category has 10000 short texts.

5.2. Experiment 1: Finding density peaks

In this experiment, we present the peaks found by the clustering method [17] of category ‘traffic’ in Data set 1, as shown in Figure 2. In the figure, the red and green spots are two density peaks in category ‘traffic’;

their ρ and δ value are relatively bigger than others. The spots locate in the upper left corner have big δ but small ρ, they are outliers. In the following experiments, we will use these density peaks to classify an unknown Chinese text.

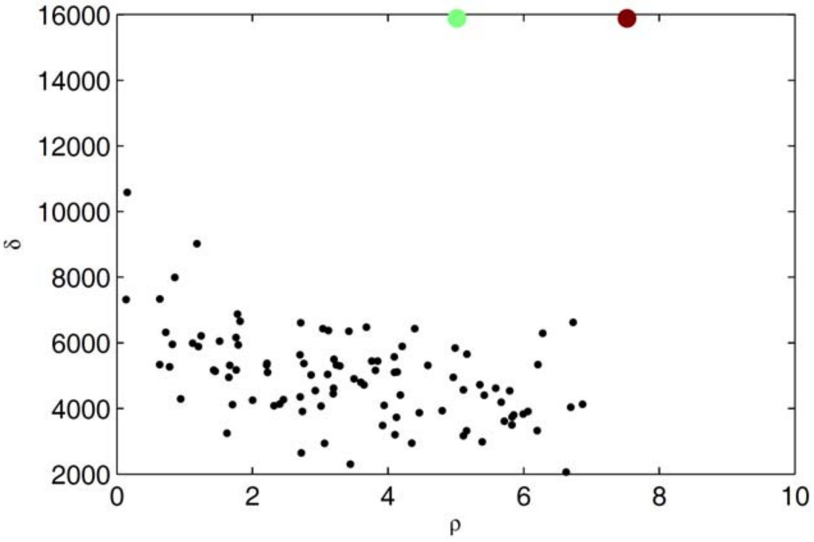


Figure 2. Peaks example of a category.

5.3. Experiment 2: Regular Chinese text

In this experiment, we test our method on Data set 1, and make comparisons with SVM, KNN. The experimental result of KNN and SVM are conducted on the tool provided by Li [15]. As mentioned above, the documents in Data set 1 are regular Chinese texts with correct and well organized usage. The experimental results with comparisons are shown the following two figures. Figure 3 shows the precision comparisons of three methods, and Figure 4 presents the recall comparisons. Clearly, we see that our method performs as well as SVM and KNN on this data set.

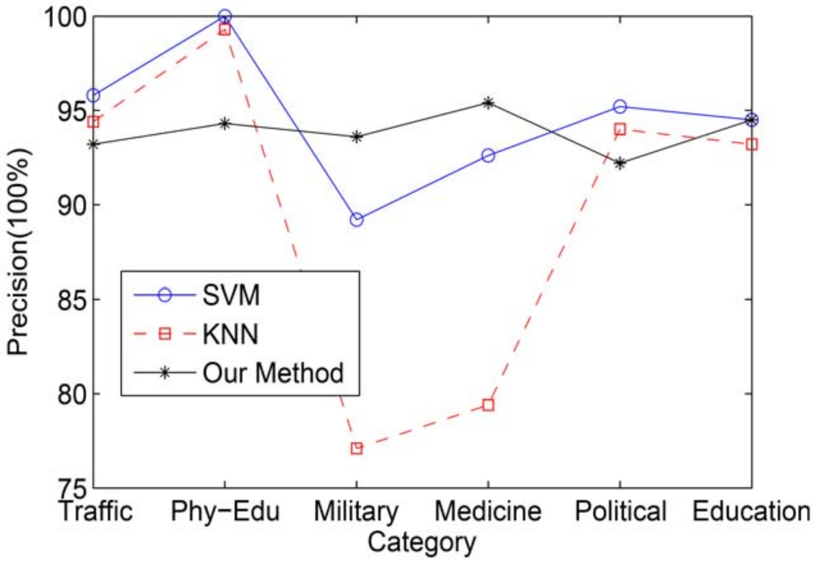


Figure 3. The comparison of precision.

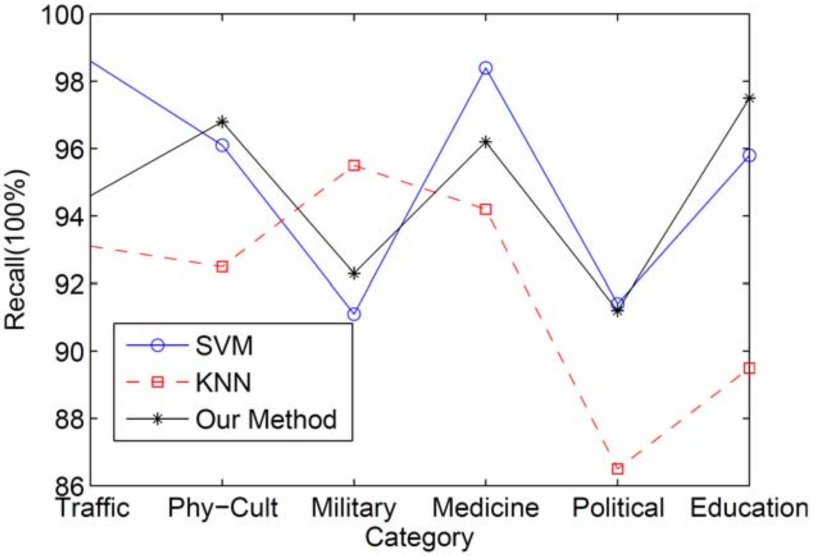


Figure 4. The comparison of recall.

Jiang et al. [12] provided LDA and labelled-LDA experimental results of F1 on Fudan NLP data set [18]. The latent aspects of labelled-LDA changes from 2 to 10, the best performance of F1 is 90.8%, while the best performance of LDA is 85.7%. In order to make a comparison, we also conduct our method on the same data set, and the results are shown in Figure 5. It is clear to show that our method achieves a better result than LDA and labelled-LDA.

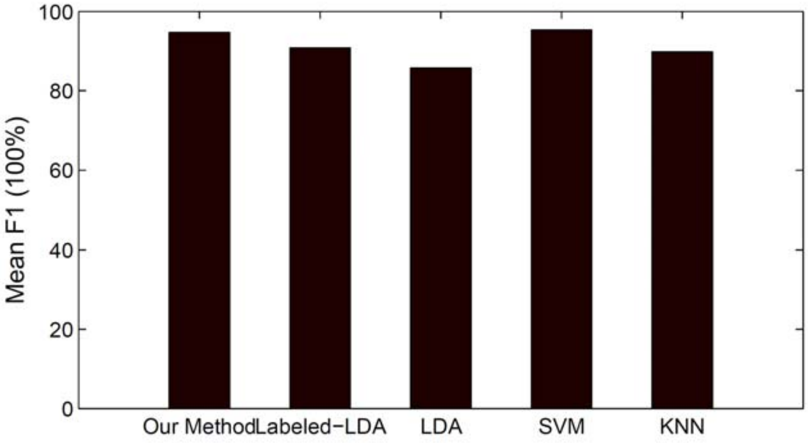


Figure 5. The comparison of mean F1.

5.4. Experiment 3: Mini training data

In this experiment, we test our method in the case of training data is not enough. It is the same as Experiment 2 that the tool of Li [15] provides the result of SVM and KNN. The mean F1 comparison is shown in Figure 6. Obviously, it shows that our method performs best in this case, and the performance is still close to the result in Experiment 2. While the performances of SVM, KNN fall rapidly, for both of them need large scale of training data.

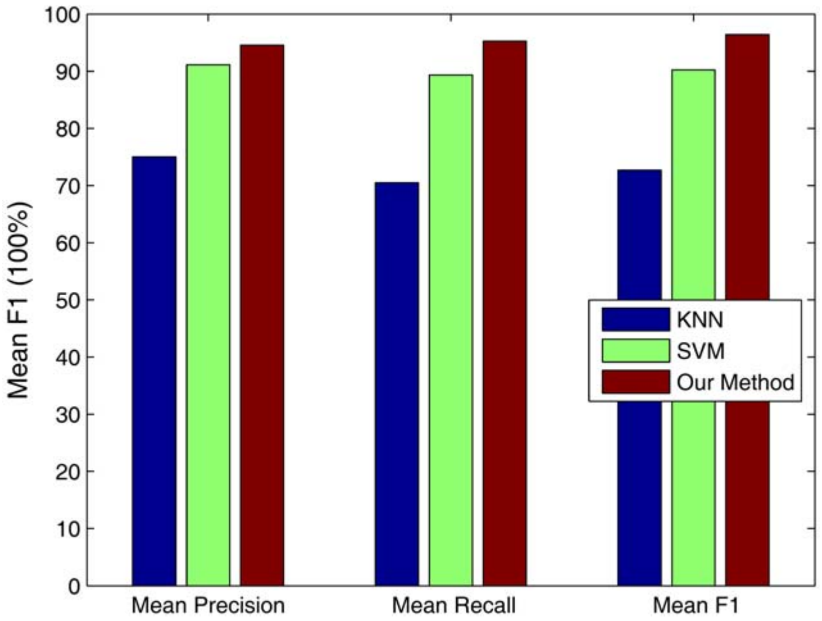


Figure 6. The comparison on mini training data.

5.5. Experiment 4: Web short texts

In this experiment, we conduct experiments on Data set 3 (SogouC), in order to evaluate the performance of our method on Chinese Web short texts. As mentioned in the previous section, many Web texts are sparse, less topic-focused, and much short, result in not enough shared context provided. Furthermore, there are many Web catchwords that are difficult to detect by segmentation tool, e.g., ICTCLAS, in these texts. Therefore, existing methods can not deal with them well. Table 2 presents the experimental result of our method conducted on Data set 3 with comparisons to CRF and SVM. The experimental data and result of CRF and SVM come from [14]; in [14], the author take 70% and 30% of the total data as training data, respectively. The data in Table 2 shows the results of SVM are not as good as the Experiment 2 and Experiment 3, but our method still works better even if our training data is much fewer than SVM and CRFs.

Table 2. The comparison on Web short texts

Precision	SVM	CRFs	Our Method
70% Training data	81.7%	88.4%	/
30% Training data	81.7%	88.4%	/
5% Training data	/	/	87.6%

6. Conclusion

The most existing machine learning methods are difficult to deal with Chinese text for some reasons, the first is the basic unit for Chinese to express is not hanzi, but Chinese phrase; the second is there is no natural delimiter to depart these Chinese phrases in Chinese text; and the third one is that Chinese phrases are variable and alternative, they are superficial expression of the real meaning. For a given semantic topic, there are many, sometimes are infinite, Chinese phrases that are semantic relevance to it. Therefore, it is impossible for a limited training data contains all of them. In the case of processing Chinese Web short texts, things are even worse; because many of these texts are short, sparse, irregular and less topic focused, as well as there are many catchwords or novel Chinese phrases in these Web texts, which adversely affect these methods, and these methods fail to achieve desire accuracy.

In order to overcome these problems, we propose a classification method for Chinese text, which makes use of BaiduBaike to extract the semantic topics of a Chinese text, and then a clustering method [17] is applied to detect density peaks of a category; then the distance between the Chinese text and each category is calculated and used to classify the text. Our experiments show that this method performs stable and well on different data set. There is not too much difference between our method and other methods in the case of the Chinese texts are regular. When the data are Web short texts or training data are not enough, the performances of other methods decrease remarkably, while our method works still well.

Acknowledgement

Supported by the grant of the National Science Foundation of China (No. 61175121); the grant of the National Science Foundation of Fujian Province (No. 2013J06014); the Promotion Program for Young and Middle-aged Teacher in Science and Technology Research of Huaqiao University (No. Z-QNYX108); the Science and Technology project of Xiamen, Fujian Province, China (No. 3502Z20133029); the Fundamental Research Funds for the Central Universities (No. JB-ZR1217); and the Natural Science Foundation of Fujian Province, China (No. 2012J05117).

References

[1] F. Sebastiani, Machine learning in automated text categorization, ACM Computing Surveys 34(1) (2002), 1-47.

[2] Jin-Shu Su, Bo-Feng Zhang and Xin Xu, Advances in machine learning based text categorization, Journal of Software 17(9) (2006), 1848-1859.

[3] Yewang Chen, Hua-Zhen Wang, Haibo Li, Binengm Zhong, Jin Gou and Duansheng Chen, A topic extraction method for Chinese web text based on BaiduBaike and text classification, Journal of Chinese Computer Systems 33(12) (2012), 2605-2610.

[4] Thomas Hofmann, Probabilistic latent semantic indexing, Proceedings of the Twenty-Second Annual, International SIGIR Conference on Research and Development in Information Retrieval (SIGIR-99), 1999.

[5] D. Blei, A. Ng and M. Jordan, Latent Dirichlet allocation, Journal of Machine Learning Research 3 (2003), 993-1022.

[6] Ge Xu and Hou-Feng Wang, The development of topic models in natural language processing, Chinese Journal of Computers 34(8) (2011), 1423-1436.

[7] C. Huang and H. Zhao, Which is essential for Chinese word segmentation character versus word, In Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation (PACLIC20), (2006), 1-12.

[8] C. Huang and H. Zhao, Chinese word segmentation: A decade review, Journal of Chinese Information Processing 21(3) (2007), 8-18.

[9] Hua-Ping Zhang, Hong-Kui Yu, De-Yi Xiong and Qun Liu, HHMM-based Chinese Lexical Analyzer ICTCLAS, Second SIGHAN workshop affiliated with 41th ACL; Sapporo Japan, (2003), 184-187.

[10] Yun-Qing Xia, Kam-Fai Wong and Pu. Zhang, Toward anomalous and dynamic nature of the Chinese network chat language, Journal of Chinese Information Processing 21(3) (2007), 83-91.

[11] Shengli Song, Shaolong Wang and Ping Chen, Chinese text semantic representation for text classification, Journal of Xidian University 40(2) (2013), 89-97.

[12] Yu-Yan Jiang, Ping Li and Qing Wang, An improved labeled latent Dirichlet allocation model for multi-label classification, Journal of Nanjing University: Nat. Sci. Ed. 49(4) (2013), 425-432.

[13] Wen-Bo Li, Le Sun and Da-Kun Zhang, Text classification based on labeled-LDA model, Chinese Journal of Computers 31(4) (2008), 621-627.

[14] Shaohua Teng, Study on Chinese Short-Text Classification, Master Degree Thesis of Tsinghua University, 2009.

[15] Ronglu Li [OL]: http://download.csdn.net/detail/superyangtze/2710559.

[16] Wikipedia [OL]: http://en.wikipedia.org/wiki/Suffix tree.

[17] Alex Rodriguez and Alessandro Laio, Clustering by fast search and find of density peaks, Science 344 (2014), 1492-1496.

[18] Fudan NLP [OL]: http://www.datatang.com/data/44082.

[19] SogouC [OL]: http://www.sogou.com/labs/dl/c.html.

■