J Ambient Intell Human Comput (2015) 6:473–480
DOI 10.1007/s12652-015-0257-z

ORIGINAL RESEARCH

A method for Chinese text classification based on apparent semantics and latent aspects

Ye-Wang Chen · Jiong-Liang Wang · Yi-Qiao Cai · Ji-Xiang Du

Received: 24 September 2014 / Accepted: 22 January 2015 / Published online: 12 February 2015

Abstract The existing methods for text classification fail to achieve high accuracy in processing Chinese texts, for that the basic unit of Chinese texts is not hanzis but Chinese phrases, and there is no natural delimiter in Chinese texts to separate the phrases. Things go even worse in the case of processing large number of Chinese Web texts, for these texts often lack of enough context, because most of these text are often short, irregular and sparse. In this paper, a new classification method is proposed for Chinese texts based on apparent semantics and latent aspects (ASLA). First, the apparent semantics of Chinese text are extracted as features instead of hanzis by BaiduBaike; Second, pLSA is applied for mining the latent aspects of these apparent semantics. Third, the relevant degree of a document to a category is calculated according to the apparent semantics and latent aspects. Finally, the category of a document is determined by the relevant degree. The proposed method is able to process Chinese web short text well with mini train data. Our experiments showed that the proposed method is promising, and it outperforms pLSA,SVM, KNN and CRF in the case of training data is not enough and the text is irregular.

Keywords BaiduBaike · Apparent semantics · Latent aspects · Chinese text classification

Y.-W. Chen · J.-L. Wang · Y.-Q. Cai · J.-X. Du (✉)
College of Computer Science and Technology of Huaqiao University Xiamen, Xiamen, China
e-mail: jxdu77@gmail.com

Y.-W. Chen
e-mail: ywchen@hqu.edu.cn

1 Introduction

There are tremendous text data emerge in the web every moment, however, most of them have nothing to do with web users, result in a dilemma that it is difficult to retrieve useful data for users. Processing and making use of these resources effectively becomes increasingly important in many Web applications. Therefore, Information Retrieval(IR) technology has been developed rapidly, and it attracts more and more people, since it provides proper data for web users. Learning to classify texts and web documents is one of the hotspots and key techniques in IR, and it has been intensively studied. Many machine learning methods discussed in (Fabrizio 2002; Su et al. 2005). Such as kNN, Naive Bayes, maximum entropy, and support vector machines, i.e. SVM, have been applied to a lot of classification problems and achieved satisfactory results.

However, it is difficult to process Chinese text, because the basic unit for Chinese to express is not hanzis, but Chinese phrase, and there is no natural delimiter in Chinese text to depart these Chinese phrases. In the Chinese web, the case is more complicated. There are many novel Chinese phrases coming out every once in a while. Normal machine learning methods can not deal with them, for they do not appear in the training data. Therefore, taking hanzis or Chinese phrases as features, these normal machine learning methods usually fail to achieve desire accuracy. We proposed some opinions about the key reasons of these failures in (Chen and Wang 2012) and (Chen and Du 2014), (1) hanzis and Chinese phrases are only the superficial part of the real semantic. Given a semantic topic, there are many, sometimes are infinite, Chinese phrases that are related to it, the training data can not contain all of them; (2) It is impossible to predict the appearance of a new Chinese phrases, and difficult to update the training data in time when a new Chinese phrase appears.

In order to overcome these problems, the real semantics of text must be extracted and treated as features, instead of hanzis or Chinese phrases. This is because the real semantics of a text express the real meaning, and they are more stable than hanzi or Chinese phrase. Starting from this idea, we propose a method for Chinese text classification. This method utilizes BaiduBaike to extract the apparent semantics of a text, and applies pLSA introduced in (Hofmann 1999) to detect the latent aspects, and then the apparent semantics and latent aspects are regarded as features to train and classify Chinese texts. Our experiments have demonstrated that this method is promising.

The rest of the paper is organized as follows. In Sect. 2 the related works are introduced. The principle of our method is presented in Sect. 3. Section 4 shows our experiments and analysis; The final section is our conclusion.

2 Related works

In recent years, there have been extensive studies and rapid progresses in automatic text classification or categorization, which is one of the hotspots and key techniques in the information retrieval and data mining field. Sebastiani (Fabrizio 2002) discussed the main approaches to text categorization that fall within the machine learning paradigm, and discussed in detail issues pertaining to document representation, classifier construction, and classifier evaluation.

So far, machine learning methods are one of the most popular techniques to classify text, such as SVM, KNN. In recent years, topic modeling, such as pLSA introduced in (Hofmann 1999) and LDA introduced in (Blei et al. 2003), plays an important role, it provides machine learning methods for automatically organizing, understanding, searching, and summarizing large electronic archives. The idea is to create a probabilistic generative model for the text documents in the corpus. These models are suitable for uncovering the hidden thematic structure in document collections, and help us develop new ways to search, browse and summarize large archives of texts. Su et al. (2005) presented a survey on the up-to-date development in text categorization based on machine learning, including model, algorithm and evaluation. Fu et al. (2015) presented a new and realistic problem, open categorical text classification, which requires to classify documents without the categorization system known beforehand. Bharti and Singh (2015) used artificial bee colony algorithm (ABC) to select appropriate cluster centers for creating clusters of the text documents. Durao and Dolog (2014) analyzed the potential of wiki technology as a tool for knowledge sharing in corporate wikis. Also there are such kind of machine learning methods used in other fields (Li et al. 2006; Li and Wang 2006).

Xu and Wang (2011) introduced the fundamental works in the development of topic models, such as LSI, pLSI and LDA, with focus on the relationship among these works, and made a simple categorization on topic models derived from LDA. Also representative models of each category were introduced. pLSA used a latent variable model that represents documents as mixtures of aspects, each aspect is represented by a distribution of words. It also associates each observation (w, d) with these aspects. In contrast to pLSA, LDA treats the multinomial weights over aspects as latent random variables. The pLSA model is extended by sampling those weights from a Dirichlet distribution, which is the conjugate prior to the multinomial distribution. This extension allows the model to assign probabilities to data outside the training corpus and uses fewer parameters, thus reducing overfitting.

In area of the Chinese text, however, it is more difficult for machine learning methods to understand and process Chinese text, for there is no natural delimiter in Chinese text and the basic unit is not hanzi, but Chinese phrases. Therefore, Chinese word segmentation is an important step in processing Chinese text, it is also an active area that has drawn great of attention in the Chinese language processing community in the past 20 years. Many technologies have been proposed, word-based method played the dominant role in the early work. Character-based tagging method in Huang and Zhao (2006, 2007) became the most popular method for its remarkable effect. ICTCLAS in (Zhang et al. 2003) provided a framework for Chinese word segmentation by HMM model, it splits regular Chinese text well. However, the result is not good if the text were Chinese web short text, e.g. the item 1 in Fig. 1, and the result is the item 2. Obviously, it is not satisfiable. In order to deal with these Chinese web short texts, Xia et al. (2007) developed a source channel model to convert short text, especially web chat text, to standard language, and

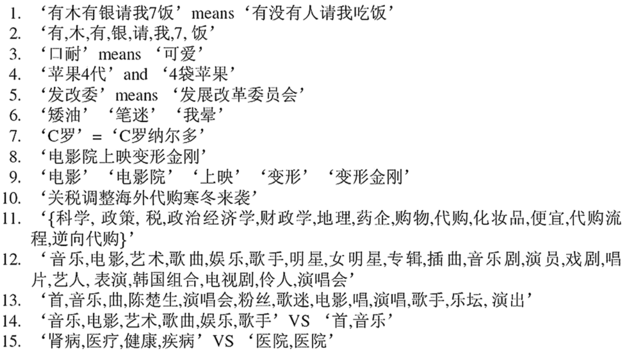

1. ‘有木有银请我7饭’ means ‘有没有人请我吃饭’
2. ‘有,木,有,银,请,我,7, 饭’
3. ‘口耐’ means ‘可爱’
4. ‘苹果4代’ and ‘4袋苹果’
5. ‘发改委’ means ‘发展改革委员会’
6. ‘矮油’ ‘笔迷’ ‘我晕’
7. ‘C罗’ = ‘C罗纳尔多’
8. ‘电影院上映变形金刚’
9. ‘电影’ ‘电影院’ ‘上映’ ‘变形’ ‘变形金刚’
10. ‘关税调整海外代购寒冬来袭’
11. ‘{科学, 政策, 税,政治经济学,财政学,地理,药企,购物,代购,化妆品,便宜,代购流程,逆向代购}’
12. ‘音乐,电影,艺术,歌曲,娱乐,歌手,明星,女明星,专辑,插曲,音乐剧,演员,戏剧,唱片,艺人, 表演,韩国组合,电视剧,伶人,演唱会’
13. ‘首,音乐,曲,陈楚生,演唱会,粉丝,歌迷,电影,唱,演唱,歌手,乐坛, 演出’
14. ‘音乐,电影,艺术,歌曲,娱乐,歌手’ VS ‘首,音乐’
15. ‘肾病,医疗,健康,疾病’ VS ‘医院,医院’

Fig. 1 Some Chinese examples

introduced phonetic mapping model constructed with standard language corpus to the source channel mode1. e.g. item 3 in Fig. 1. Chen et al. (2014) pointed out that it is necessary to construct a particular lexical database for Cantonese, and explored some Cantonese special written-tradition rules and incorporate them into the feature-based opinion summarization system framework.

Also there are many methods for Chinese text classification. Jiang et al. (2013) proposed an improved Labeled-LDA model for multi-label classification. In this model, labels have two components which are local topics and shared topics. The prediction of label is a combination of local topics and shared topics. Song et al. (2013) proposed a new Chinese text semantic representation model by considering contextual semantic and background information on the words with wikipedia. The model retained the contextual information on each word with a large extent. Teng (2009) introduced a method based on CRFs for web short text, character-based and character-tagging were used to extract feature. Li et al. (2008) proposed another Labeled-LDA method to enhance the traditional LDA to integrate the class information, and a new algorithm was introduced to determine the quantities of latent topics for each class. Li (2010) provided a tool for classifying Chinese texts based on SVM and KNN, the ICTCLAS was used for Chinese word segmentation.

These machine learning methods work well on the regular texts. However, there are two shortages, the first is great mounts of training data is needed; the second is in the case of texts are short, sparse, irregular and less topic focused, furthermore, there are many catchwords or novel Chinese phrases in these web text, these methods fail to achieve desire accuracy.

3 The principle

3.1 The insufficient of existing methods

Usually, word segmentation is an important step for processing Chinese text of word-based method, another important step is to attain the statistical feature of the segmentation result, and then fulfill the task of training and classification by different mathematic model. However, the result is severely affected by the Chinese words segmentation. Take item 4 in Fig. 1 for example, both of the two sentences contains one same Chinese phrase, while the real mean of them are totally different. For another example, there are many shorthand phrases in Chinese text, such as item 5 in Fig. 1, the similarity between the two strings is zero. The existing methods fail to deal with these cases.

We argue that the Chinese words or phrases are only the manifestation of the real semantic. For a given semantic topic, there are many, sometimes are infinite, Chinese phrases or words that are related to it. Therefore, it is impossible to enumerate all Chinese phrases in training data. In another words, Chinese phrases or words are fickle and unreliable. While, the semantic topic and the topic relation of these endless Chinese phrases or words are stable, such as 'F35' 'F22' ... are all a kind of 'Battle Plane' (means Battle Plane). Obviously, 'Battle Plane' is a stable and abstract concept that are related to military affairs, and this concept indicates the connotation that expresses the high hierarchical relation of this area. According to 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 text with stable relationship among these semantics.

In Chinese web, BaiduBaike is an open and free knowledge base, it provides comprehensive, accurate and complex information about a Chinese phrase. Furthermore, it keeps up with the hot spots and network catchwords. There are some superiorities of BaiduBaike as the following, (1) comprehensiveness: there are about 3.4 million BaikePhrases in BaiduBaike so far. General speaking, it covers all domains of the society, even network catchword, such as item 6 in Fig. 1. (2) Realtime: 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 item 7 in Fig. 1.

Many of these Baike-Phrases, e.g. item 6 in Fig. 1, are not identified by the current Chinese segmentation tools, such as ICTCLAS, which makes adverse effects on understanding the text for the machine learning methods, and then leads to many mis-classifications for Chinese web text inevitably. Therefore, it is feasible for Baidu-Baike to be an infrastructure that provides accurate, real time and rich information for Chinese text mining and classification.

3.2 Baike-phrase, semantic topic, Chinese text

The basic unit of Baidu-Baike is Baike-Phrase, which is composed of 6 parts: signature, main body, reference, open class, related phrases and extension. Every part describes the Baike-Phrase from different aspect. We note that the "open class" presents the deep semantic in the hierarchial knowledge of a Baike-Pharse, in another words, it is a stable and abstract concept that expresses the connotation of the Baike-Pharse. Therefore, it is useful for us to extract the semantics of Chinese text.

Definition 1 Apparent sematic topic: an open class of a Baike-Phrase is an apparent semantic topic.

There are some basic points of view in (Chen and Wang 2012) by our observation and analysis 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. Apparent 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 an apparent semantic topic in a Chinese text, the more Baike-Phrases related to it there are; Two Chinese texts with similar semantics have similar apparent semantic topics.

3.3 The basic idea

According to the points of view above, we have the idea that a Chinese text can be mapped into a set of apparent semantic topics from a set of Baike-Phrases, then LDA or pLSA is used to catch the latent aspects of these apparent semantic topics, and these apparent semantics and latent aspects can be treated as features for classifying Chinese text. By this way we avoid facing Chinese phrases or words directly, and solve two problems of the existing methods to a great extent, the first is great mounts of training data are needed, and second is web short texts are difficult to process.

Therefore, in this paper we propose a method for Chinese text classification, it requires less training data than existing method, and performs well on not only regular Chinese text but also web short text. The main processes are the following. (1) Detect all possible Baike-Phrases in a Chinese text; (2) Extract all apparent semantic topics of the text; (3) Mining the latent aspects of these apparent semantic topics by pLSA for each category; (4) The latent aspects and apparent semantic topis are used as features for our classification model; (5) Classify a text by relevant degree calculated in the model.

4 The processes

4.1 Detect baike-phrases in a Chinese text

Definition 2 Candidate-phrase: given C is a text, $C_{i,j}$ is a string that starts from *ith* char to *jth* char of C, $C_{i,j}$ is a Candidate-Phrase if $C_{i,j}$ was a Baike-Phrase.

We build a prefix base (Sartaj 1999) for all Baike-Phrases firstly, and then it is easy to detect all Candidate-Phrases effectively, the algorithm was presented in our previous work (Chen and Wang 2012). Take item 8 in Fig. 1 for example, there are 4 Candidate-Phrases, i.e. item 9 in the same figure.

4.2 Extract apparent semantic topics

Definition 3 Semantic relevance: Let e be 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 apparent 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.

For a Chinese text document, we extract all relevant apparent semantic topics of it and put them into a topic set. Take item 10 in Fig. 1 for example, the topic set is item 11. The semantic topics in the set reveal the deep mean 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.

4.3 Mining latent aspects

In this section, we will focus on using probabilistic latent semantic indexing (pLSA) method. For pLSA, a document is regarded as a mixture of underlying (latent) K aspects (or latent semantic), each aspect is represented by a distribution of words $P(w|z)$. It associates each observation (w, d) with a latent variable $z \in Z\{z_1, \ldots, z_K\}$. It selects a document with probability $P(d)$, picks a latent variable z with probability $P(z|d)$, generates a word w with probability $P(w|z)$.

In our method, we replace the w in pLSA with e, i.e. apparent semantic topic, and use pLSA to determine the latent aspects of apparent semantic topics with their probabilities, i.e. $P(z)$. $P(z)$ is treated as the weight of aspect z, and $P(e|z)$ is the weight of e that belongs to aspect z. And then we build a log-likelihood function as follows.

$$L = \sum_{d \in D} \sum_{e \in E} n(d,e) log P(d,e) \tag{1}$$

where E is the apparent semantic topics set of all documents, D is the document collection.

In order to maximize it, EM in (Dempster et al. 1977) is used for this purpose. The probability that a apparent semantic topic e occurs in a document d, is explained by aspect z.

$$P(z|d,e) = \frac{P(z)P(d|z)P(e|z)}{\sum_{z'} P(z')P(d|z')P(e|z')} \tag{2}$$

and then,

$$P(e|z) = \frac{\sum_{d} n(d,e)P(z|d,e)}{\sum_{d,e'} n(d,e')P(z|d,e')} \tag{3}$$

$$P(d|z) = \frac{\sum_{e} n(d,e)P(z|d,e)}{\sum_{d',e} n(d',w)P(z|d',e)} \tag{4}$$

$$P(z) = \frac{\sum_{d,e} n(d,e)P(z|d,e)}{\sum_{d,e} n(d,e)} \tag{5}$$

Next, for each category $c \in C\{c_1, c_2, \ldots c_L\}$, we train the training corpus, and then we have $P^{(c)}(z)$, and $P^{(c)}(e|z)$ for each e and z in category c, respectively.

4.4 Text classification model

In order to determine the category of a new document d, i.e. classify it, some steps are involved in our method as the following. Firstly we extract all candidate-phrases of d and form them as a set CP_d; Secondly, all relevant semantic topics of document d are extracted, and then count the appearance number of each $e \in E\{e_1, e_2, \ldots, e_n\}$ in document d, noted as $num_d(e)$

$$num_d(e) = \sum_{cp \in CP_d} \sigma(i, cp) \tag{6}$$

where $\sigma(i, cp) = 1$ if e_i is an apparent semantic topic of Baike-Phrase cp, or else $\sigma(i, cp) = 0$.

And then determine the relevant degree (RelDeg) between the new document d and a category of c as follows.

$$RelDeg^{(c)}(d) = \sum_{z \in Z} P^{(c)}(z)\left[\sum_{e \in E} P^{(c)}(e|z) * num_d(e)\right] \tag{7}$$

Finally, the goal to classify document d is to find a category c that maximize the relevant degree, i.e.

$$arg \max_{c \in C} RelDeg^{(c)}(d) \tag{8}$$

5 Experiments and analysis

In this section, we conduct a series of experiments for evaluating the proposed ASLA, and carry out comparisons with KNN, SVM, LDA, pLSA and CRF. The experimental platform is Intel Centrino Duo T2400 1.83 GHz + 3G Memory + WindowsXP SP2.

5.1 Experimental data

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

Data set 1 is a regular Chinese documents collection that downloads from (FudanNLP 2013), 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 2013) (Full version), and we only extract the title of each document as web short text. we also select 6 categories, each category has 10,000 short texts.

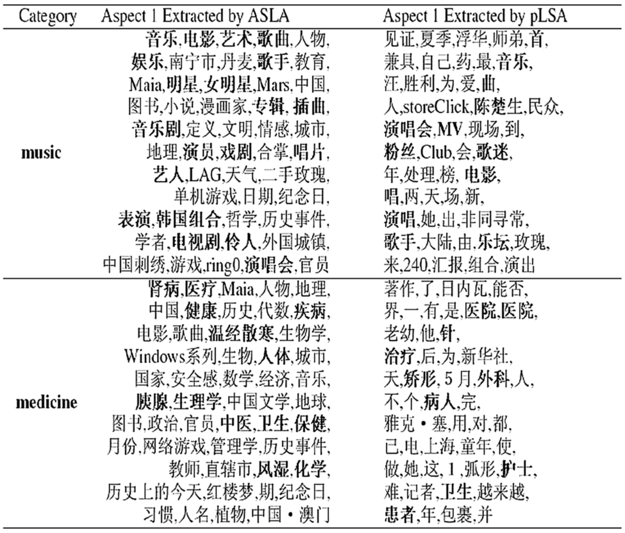
5.2 Experiment 1: latent aspect extracted by our method

This section shows the experiments of extracting the first latent aspects of two categories texts, i.e. *music* and *medicine*, by the proposed method ASLA and pLSA, respectively. The experimental data come from (SogouC 2013), and the results are shown in Fig. 2. Those Chinese phrases listed in the table are the extracted result that belong to the first aspect with top 50 probability.

As the table shows, obviously in the “music” category, the phrases in the first column are better than the second column, for there are 20 Chinese phrases that relevant to “music” directly with high probabilities. These Chinese phrases are bold in the table, i.e. item 12 in Fig. 1. While there are only 13 directly relevant phrases extracted by pLSA, i.e. item 13. It is clearer to see that ALSA is superior to pLSA in the case of focusing on the top 10 Chinese phrases only, the pLSA is defeated by 6-2, i.e. item 14. In the “medicine” category, we also see that the ALSA is superior to pLSA, for there are more Chinese phrases that are relevant to “medicine” in the first column. Also, it is more obvious to see that ALSA is better in top 10 Chinese phrases, the pLSA is defeated by 4-2, i.e. item 15.

Table 1 The description of training and test data

Train/test	Data set 1	Data set 2	Data set 3	
Traffic	2,000/715	30/715	Women	500/9,500
Phy-Edu	1,000/482	30/482	Phy-Edu	500/9,500
Military	2,000/435	30/435	Military	500/9,500
Medicine	1,500/543	30/543	News	500/9,500
Politics	2,500/701	30/701	Travel	500/9,500
Education	1,500/550	30/550	Education	500/9,500



Category	Aspect 1 Extracted by ASLA	Aspect 1 Extracted by pLSA
music	**音乐**,**电影**,**艺术**,**歌曲**,人物,	见证,夏季,浮华,师弟,首,
	娱乐,南宁市,丹麦,**歌手**,教育,	兼具,自己,药,最,**音乐**,
	Maia,**明星**,**女明星**,Mars,中国,	汪,胜利,为,爱,**曲**,
	图书,小说,漫画家,**专辑**, **插曲**,	人,storeClick,**陈楚生**,民众,
	音乐剧,定义,文明,情感,城市,	**演唱会**,**MV**,现场,到,
	地理,**演员**,**戏剧**,合掌,**唱片**,	**粉丝**,Club,会,**歌迷**,
	艺人,LAG,天气,二手玫瑰,	年,处理,榜, **电影**,
	单机游戏,日期,纪念日,	唱,两,天,场,新,
	表演,**韩国组合**,哲学,历史事件,	**演唱**,她,出,非同寻常,
	学者,**电视剧**,**伶人**,外国城镇,	**歌手**,大陆,由,**乐坛**,玫瑰,
	中国刺绣,游戏,ring0,**演唱会**,官员	来,240,汇报,组合,演出
medicine	**肾病**,**医疗**,Maia,人物,地理,	著作,了,日内瓦,能否,
	中国,**健康**,历史,代数,**疾病**,	界,一,有,是,**医院**,**医院**,
	电影,歌曲,**温经散寒**,生物学,	老幼,他,**针**,
	Windows系列,生物,**人体**,城市,	**治疗**,后,为,新华社,
	国家,安全感,数学,经济,音乐,	天,**矫形**,5月,**外科**,人,
	胰腺,**生理学**,中国文学,地球,	不,个,**病人**,完,
	图书,政治,官员,**中医**,**卫生**,**保健**,	雅克・塞,用,对,都,
	月份,网络游戏,管理学,历史事件,	已,电,上海,童年,使,
	教师,直辖市,**风湿**,**化学**,	做,她,这,1,弧形,**护士**,
	历史上的今天,红楼梦,期,纪念日,	难,记者,**卫生**,越来越,
	习惯,人名,植物,中国・澳门	**患者**,年,包裹,并

Fig. 2 The extraction of the first latent aspects of two categories texts by ALSA and pLSA

5.3 Experiment 2: regular Chinese text

In this part, we implement our method ASLA, and the method of PLSA (pLSA+ICTCLAS) for text classification. Then we conduct an experiment on Data set 1, and make a comparison with SVM, KNN and PLSA. The test result of KNN and SVM comes from the tool of Li (2010). The configuration of SVM and KNN is the following. Feature selection is information gain, language is Chinese, word segmentation tool is ICTCLAS, feature dimension is 1000.

As mentioned above, the content of the documents in Data set 1 are regular Chinese texts, with correct usage of Chinese phrases and well organized. The comparisons are shown in Figs. 3 and 4, we can see that our method ASLA performs well on regular Chinese texts, that is close to SVM, and superior to KNN and PLSA.

Jiang et al. (2013) conducted experiments on FudanNLP data set and provided mean F1 of Label-LDA and LDA, we compare our method ASLA with them on the same data set, the result is shown in Table 2. The latent aspects of Labeled-LDA changes from 2 to 10, the performance of F1 varies in the range of (88, 90.8 %), while LDA varies between 84.2 and 85.7 %. Obviously, it shows that with the help of apparent semantics and latent aspects, ASLA arrives at better result than LDA.

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 (2010) provides the result of SVM and KNN. The mean F1 comparison is shown in Fig. 5. Obviously, it shows that our method outperforms the others, and the performance is still close to the result in experiment 2. While the performances of SVM, KNN and PLSA fall rapidly, for all of them need great mounts of training data. From this experiment, we can see that the proposed method ASLA can performs with few train data.

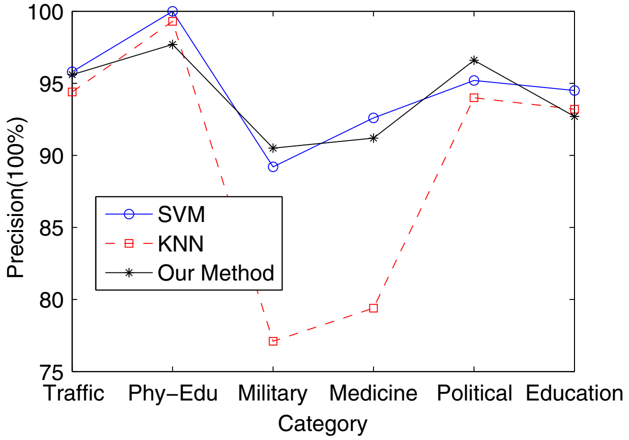


Fig. 3 The comparison of precision

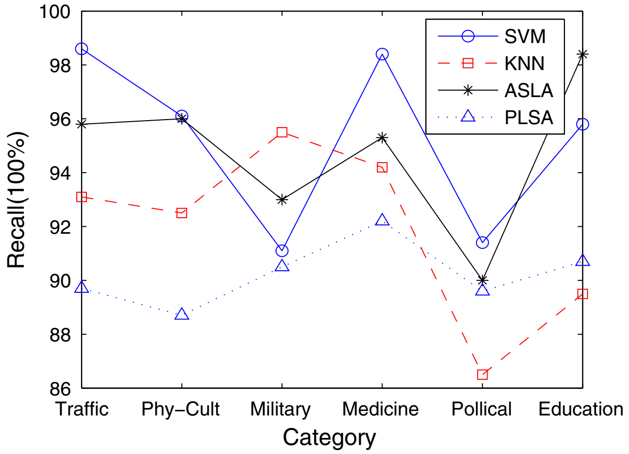


Fig. 4 The comparison of recall

Table 2 The comparison of mean F1

ASLA (%)	Labeled-LDA (%)	LDA (%)	SVM (%)	KNN (%)
93.7	88–90.8	84.2–85.7	95.3	89.8

5.5 Experiment 4: web short text

This experiment is for verifying our method on Chinese web short texts. Many web texts are sparse, less topic focused, and much short, for these texts often do not provide enough words or shared context. Furthermore, there are

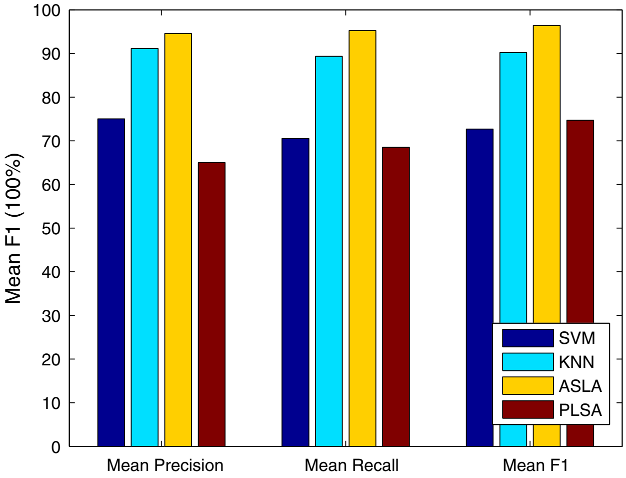


Fig. 5 The comparison on mini training data

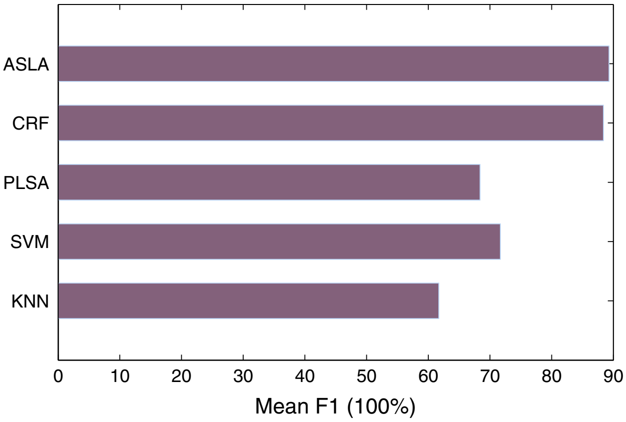


Fig. 6 The comparison on web short text

many network catchwords that are difficult to detect by segmentation tool, e.g. ICTCLAS, in these texts. Therefore, existing methods can not deal with them well. Figure 6 presents mean F1 of our method conducted on data set 3 with comparison to CRF, kNN, SVM, and PLSA. where the result of CRF comes from (Teng 2009). Figure 6 shows the results of kNN , SVM and PLSA are not as good as the experiment 2 and experiment 3, but ASLA still works well, and has a slight advantage to CRF. From this experiment, we can see that the proposed method ASLA has the ability to deal with Web short text well.

6 Conclusion

6.1 Our contribution

It is difficult for machine learning methods to process Chinese text, because the basic unit for Chinese to express is not hanzi, but Chinese phrase. And there is no natural delimiter in Chinese text to depart these Chinese phrases. Furthermore, Chinese phrases are only the superficial part 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. It is even worse for these methods to process Chinese web texts, because many of these texts are short, sparse, irregular and less topic focused. Also there are many catchwords or novel Chinese phrases in these web texts, which adversely affect these methods, these methods fail to achieve desire accuracy.

In order to overcome these problems, a method for classifying Chinese text is proposed, it makes use of BaiduBaike to extract the apparent semantic topics of Chinese text, and then pLSA is applied to detect the latent aspects of a category. Then this method treats them as features in our model to train and classify. 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 better.

6.2 Future works

There are many typing mistakes and shortening expressions in the web, which make great effect on our method. Therefore, our further work will be carried out in using string matching and Phonetic-Based approach (Durao and Dolog 2014) to improve our method.

Acknowledgments Supported by the Grant of the National Science Foundation of China (No. 61175121); the Grant of the National Science Foundation of Fujian Province (No. 201-3J06014); the Promotion Program for Young and Middle-aged Teacher in Science and Technology Research of Huaqiao University (No. ZQNYX108); the Fundamental Research Funds for the Central Universities (No. JBZR-1217); the Natural Science Foundation of Fujian Province, China (No. 2012J05117).

References

Bharti KK, Singh PK (2015) Chaotic gradient artificial bee colony for text clustering. Soft Comput. doi:10.1007/s12652-014-0237-8

Blei D, Ng A, Jordan M (2003) Latent dirichlet allocation. J Mach Learn Res 3(1):993–1022

Chen J, Huang DP, Hu SY, Liu Y (2014) Cai Y, Min HQ An opinion mining framework for cantonese reviews. J Ambient Intell Humaniz Comput. doi:10.1007/s12652-014-0237-8

Chen YW, Du JX (2014) A new method for classifying chinese text based on semantic topics and desity peaks. Int J Appl Math Mach Learn 1(1):35–54

Chen YW, Wang HZ et al (2012) A topic extraction method for chinese web text based on baidubaike and text classification. J Chin Comput Syst 33(12):2605–2010

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the em algorithm. J R Stat Soc 39(1):1–38

Durao F, Dolog P (2014) Improving tag-based recommendation with the collaborative value of wiki pages for knowledge sharing. J Ambient Intell Hum Comput 5(1):21–38

Fabrizio S (2002) Machine learning in automated text categorization. ACM Comput Surv 34(1):1–47

Fu RJ, Qin B, Liu T (2015) Open-categorical text classification based on multi-lda models. Soft Comput 19(1):29–38

Fudan NLP (2013) Chinese texts database. IOP Publishing PhysicsWeb. http://www.datatang.com/data/44082. Accessed 11 September 2014

Hofmann T(1999) Probabilistic latent semantic analysis. In: Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, Stockholm, Sweden, pp 289–296

Huang C, Zhao H (2006) Which is essential for chinese word segmentation:character versus word. In: Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation, pp 1–12

Huang C, Zhao H (2007) Chinese word segmentation: a decade review. J Chin Inf Process 21(3):8–18

Jiang YY, Li P, Wang Q (2013) An improved labeled latent dirichlet allocation model for multi-label classification. J Nanjing Univ Nat Sci Ed 49(4):425–432

Li J, Wang YM (2006) Universal designated verifier ring signature (proof) without random oracles. Emerging Dir Embed Ubiquitous Comput 4097:332–341

Li J, Zhang FG, Wang YM (2006) A new hierarchical id-based cryptosystem and cca-secure pke. Emerging Dir Embed Ubiquitous Comput, 4097:362–371

Li RL (2010) Svmcls IOP Publishing PhysicsWeb. http://download.csdn.net/detail/superyangtze/2710559. Accessed 8 Sept 2014

Li WB, Sun L, Zhang DK (2008) Text classification based on labeled-lda model. Chin J Comput 31(4):621–627

Sartaj S (1999) Data structures, algorithms, and applications in java suffix trees. IOP Publishing PhysicsWeb. http://www.cise.ufl.edu/ sahni/dsaaj/enrich/c16/suffix.html. Accessed 11 Sept 2014

SogouC (2013) Sogou lab data. IOP Publishing PhysicsWeb. http://www.sogou.com/labs/dl/c.html. Accessed 12 Sept 2014

Song SL, Wang SL, Chen P (2013) Chinese text semantic representation for text classification. J Xidian Univ 40(2):89–97

Su JS, Zhang BF, Xu X (2005) Advances in machine learning based text categorization. J Softw 17(9):1848–1859

Teng SH (2009) Study on chinese short-text classification. Master's thesis, Tsinghua University

Xia YQ, Wong KF, Zhang P (2007) Toward anomalous and dynamic nature of the chinese network chat language. J Chin Inf Process 21(3):83–91

Xu G, Wang HF (2011) The development of topic models in natural language processing. Chin J Comput 34(8):1423–136

Zhang HP, Yu HK, Xiong DY, Liu Q (2003) Hhmm-based chinese lexical analyzer ictclas. In: 2nd SIGHAN workshop affiliated with 41st ACL; Sapporo Japan, July 2003, pp 184–187