Research of specific field ultra-short text classification based on collaborative filtering algorithm

Weichen Yang*, and *Yanwei* Si

HeBei Aisino Technology Co., Ltd, Hebei, China

Abstract. In some specific fields, there are a lot of ultra-short texts that need to be categorized. This paper proposes an ultra-short text classification method based on collaborative filtering algorithm aiming at the problems such as short text content, short length, sparse features, and large number of categories in certain fields. First, converting ultra-short text into word frequency vector by doing Chinese word segmentation and calculating word frequency; Secondly, combining relevant data in specific fields, defining the ultra-short texts as users, categories as items, and then constructing a user-item recommendation matrix. Finally, calculating text similarity by using cosine similarity method and obtaining the classification results. The experimental results show that the proposed method can well solve the problem of classification of ultra-short texts in specific fields, and the average accuracy is 9.19% and 3.81% higher than vector space model and topic similarity method respectively.

1 Introduction

In recent years, with the advent of the web2.0 era, a large number of short text web data are generated on the internet[1]. There are a lot of valuable information in these web short text data. At present, the classification of these data, and how to obtain the key information from the text more quickly and accurately, have become the key issues in current data mining research.

However, in these network data, there are some ultra-short text data in certain specific fields. These ultra-short texts consist of only a few phrases. For example, the length of a product name in a tax invoice is often within 20 words. How to dig out the target information from these phrases becomes a new challenge. The vector will be sparse when embedding a phrase using the traditional vector space model. Especially in the test phase, due to insufficient training data features, many useful features may not be captured by the model. Therefore, using the traditional short text classification method will result in unsatisfactory classification results. Aiming at the problems such as few words, sparse characters and many kinds of categories in ultra-short texts, this paper proposes a new ultra-short text classification method that combines the special relevant data features in some

* Corresponding author: yangwshbhx@163.com

specific fields and recommendation algorithms, and can classify the ultra-short texts quickly and accurately in specific fields.

The structure of this paper is as follows: section 2 introduces the research status of short text classification, section 3 describes the collaborative filtering algorithm model, and section 4 proposes a short text classification based on collaborative filtering. Section 5 is the analysis of the experiment and results, section 6 gives the summary.

2 Related works

Different from traditional texts, short text has some features such as sparseness, strong real-time, non-standard wording, and many new words[2]. Based on traditional texts such as KNN [3], Bayesian classification [4], decision tree [5], SVM [6], and maximum entropy [7], previous classification methods have achieved good results in text classification. However, these methods require enough co-occurrence information of word frequency in texts, and it is not effective when applying to short text classification.

At present, there are mainly two types of general short text classification methods. The first method based on a search engine, treats short text as a query, and expands the short text using the retrieval result from search engine. BOLLEGALA, SAHAMI et al[8-9] regarded the words contained in the short texts as queries, and input them into search engine, then used support vector machines to integrate similarity scores for measuring the semantic similarity. However, this kind of method is weak in the processing of noise information, and it relies heavily on search engine tools.

The other is a large scale corpus based approach. Wang Sheng, Fan Yunjie, Zhao Hui [10-12] used the "How-Net" to determine the upper and lower relation of the word pair, and then used the relation to extend the feature vector of the short text. Although this kind of method can obtain the correlation between semantics, but it can not be used to deal with the words not exist in the corpus.

In recent years, aiming at the short comings of the two types of short text classification method, scholars have studied the short text classification more widely. Phan et al. proposed to use the subject of short text as an additional feature set[13]; Yang Mengmeng et al. proposed a method based on LDA model topic distribution similarity classification [14]; Ma Chenglong et al. used the Gauss mixture model to carry out the general background semantic model training for unlabeled text data[15]. These methods are based on the external correlation data for feature extension or similarity estimation, and good results has been achieved. However, these methods are highly relevant to the external data, and these related data are usually domain knowledge and generally difficult to obtain in practical applications.

To sum up, these methods have made some good progress in the field of short text classification, but do not specializes in the study of ultra-short text information. In some specific fields, short texts composed of phrases often contain less features, more fresh words, and more categories to be divided. It is very difficult to classify these ultra-short texts quickly and accurately.

In our proposal, we extract some of the relevant information in specific field, take them as features of the collaborative filtering algorithm to calculate the similarity between the ultra-short text, so that implement classification of the texts.

3 Model of collaborative filtering

3.1 Collaborative filtering

Collaborative filtering algorithm is one of the earliest and most successful algorithm in recommendation systems. By analyzing the historical score data of the user on items, the method finds the similarity between the users, and then finds the similar users of the target users according to the nearest neighbor technology, and uses these similar users' evaluation of items to predict the preference of the target users to the specific items, and then to complete the recommendation of the target users. In short, the collaborative filtering approach is to find a set of users that have similar interests with the target user and then recommend the content they are interested in to the target user. The basic schematic diagram is shown in Fig.1: Since both the user A and the user B like the item A and the item B, the users A and B have similar preferences, and thus the item C that the user B likes can be recommended to the user A.

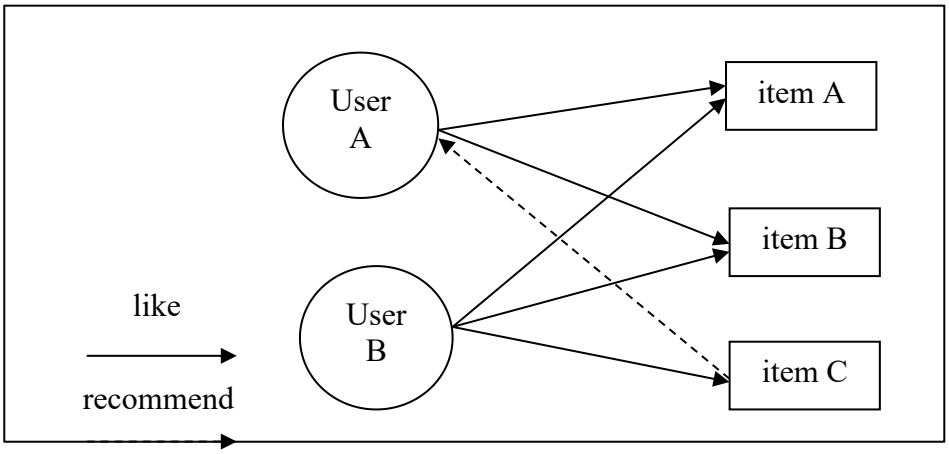

Fig. 1. Collaborative Filtering Schematic Diagram.

3.2 Cosine similarity metrics

The core of collaborative filtering recommendation method is to calculate the similarity between users. Only when similar users are found, can the recommendation be completed. In machine learning algorithms, there are various ways to measure user's distance or similarity, such as Manhattan distance, Euclidean distance, cosine similarity, etc. Among them, cosine similarity is widely used in collaborative filtering algorithms.

The cosine similarity measures the difference between two individuals using the cosine of the angle between two vectors in the vector space. Compared to other distance metrics, the cosine similarity pays more attention to the difference in the direction of the two vectors, rather than the distance or length. The formula is as follows:

$$sim(X,Y) = cos\theta = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \cdot \|\mathbf{y}\|} \tag{1}$$

According to the collaborative filtering model described above, this paper believes that the classification problem in some specific areas can be approximate described by appropriate model. Since the classification problem in short text classification can be converted into a recommendation problem , based on the collaborative filtering model, this paper proposes a mixture recommendation model based on relevant data in specific fields.

4 Ultra-short text classification based on collaborative filtering

In practical applications, short text data in some specific fields consists of only several phrases. For example: The name of the goods on the tax invoice is usually some phrases like Samsung Mobile Phone, Apple, and Sewer, etc. These ultra-short texts are not only shorter in length, but also have more fresh vocabulary. It is difficult to accurately classify these data using traditional short text classification methods. However, these specific areas often contain some special correlation data. For example, tax invoices generally provide a pre-estimation of a category by the merchant based on the product name, which is equivalent to that the merchant gives a preference category based on the item. Therefore, this article translates this kind of ultra-short text classification problem into a recommendation problem. First, the user (product name) and the item (data to be categorized) information is constructed into a recommendation matrix, then the similarity value of the short text is calculated according to the cosine similarity, and finally, the classification result is obtained in combination with the data to be classified.

4.1 Similarity calculation of ultra short text

Because the text classification model is based on collaborative filtering algorithm, cosine similarity is used to calculate the similarity between ultra short texts. First, the information to be classified of ultra short text is defined as the column, the ultra short text information is defined as the line, and the user project matrix is constructed; Then, the similarity between the users is obtained by the cosine similarity, and the user similarity matrix is constructed.

In order to find similar ultra-short texts, the cosine similarity between vectors is used to compute similar values. Suppose that training data contains a series of ultra-short text documents, $D = \{d_1, d_2, d_3, ... d_n\}$, di represents an ultra-short text, and there are n training data in all, these data need to be divided into C categories, among them, $C = \{c_1, c_2, c_3, ... c_m\}$, ci represents a category, and there are m categories in all. If an ultra-short text di is pre-classified to c_i, then the feature value of vector is recorded as 1, that means $E(U_k/P_i)=1$. When the same kind of ultra-short text is pre-classified into multiple categories, the feature values of vector are added. The calculation of the specific value is as shown in formula 2:

$$E\left(U_k / P_i\right) = \sum_{i=1}^{m} \theta_i \tag{2}$$

According to the constructed vector model, the cosine similarity between vectors is calculated, and the similarity values between ultra-short texts are obtained.

4.2 Ultra-short text classification algorithm

After getting the values of similarity between the ultra-short texts, because the similar texts will produce multiple pre-estimated classes, the KNN classification method is used to classify the pre-estimated classes. KNN classification method, also known as K-Nearest Neighbor, is a machine learning method in the field of data mining, and is one of the most important methods in classification methods. The KNN classifier has a good text classification effect. Its basic idea is to represent all text in the training set and the test set to vector form, then calculate the similarities between every text in test set and all texts in training texts, and show them in descending order, then find the most front K texts, and judge the text to be classified according to the category of the K texts. The classification methods of classifying the text according to the K texts are as follows: according to the similarity between the K texts and the text to be classified, sum the similarity of the same category, then sort the similarity of each class, and the most likely category is the final category. The similarity summation formula 3 is as follows:

$$T_k(\theta_{td}) = \sum_{k=1}^{K} \delta_k(ld_i) D_{tl}(\theta_t \mid \theta_l) \tag{3}$$

$D_{tl}(\theta_t \mid \theta_l)$ represents the similarity of text td and text ld in test set. $\delta_k(ld_i)$ represents the probability that text ld_i belongs to class k, $T_k(\theta_{td})$ represents the score that text td belongs to a certain category in test set.

5 Experiment

The experimental data is the tax invoice data provided by the tax department. A total of 10 categories are selected. The contents of each invoice are used as a data information. The name of the commodity in the tax invoice is regarded as the ultra short text data. The total data contains 94037 invoice information, of which 61685 are assisted data to be classified. The detailed data information is shown in table 1.

Table 1. detailed distribution of text quantity

class	number of ultra-short texts
Wholesale of building materials	11115
Retail of general merchandise	9312
Retail of automobile	10987
Retail of hardware	9728
Jewelry	8531
Coating manufacturing	8827
Wholesale of western medicine	9848
Medical equipment	8319
Chemical products	8256
Manufacturing products	9114

5.1 Pre-processing

In dealing with the experimental data, the Chinese Academy of Sciences segmentation tool is used to segment the ultra-short text and remove the stop word in the result of segmentation. Since ultra-short text is usually a phrase made up of several words, for each of the ultra short texts, when more than 3 words are the same as another one, the two ultra short texts are considered to be the same ultra-short text and should be merged.

5.2 Experimental evaluation

The accuracy rate (Precision), recall rate (Recall) and F1 value were used to evaluate the classification results. The identified ultra-short text category is compared with the artificial annotation. When the classification of the experiment is exactly the same as that of the artificial annotation, it is considered that the ultra short text is classified accurately. The calculation methods of accuracy, recall and F1 value are shown in formula 4, 5 and 6:

$$P_i = l_i / m_i \times 100\% \tag{4}$$

$$R_i = l_i / n_i \times 100\% \tag{5}$$

$$F1_i = 2 \times (P_i \times R_i) / (P_i + R_i) \times 100\% \tag{6}$$

l_i represents the number of ultra short texts that are classified into class i and correctly classified, and mi represents the number of ultra short texts classified into class i, and ni represents the number of ultra-short texts belong to class i.

In order to compare the effect of the proposed method with other methods, the accuracy and recall of the traditional vector space model and the topic similarity method are listed in table 2.

Table 2. Comparison of experimental results of three classification methods.

class	VSM		Topic similarity		Ultra-short classification	
	Precision	Recall	Precision	Recall	Precision	Recall
Building materials	68.34%	66.87%	74.43%	71.76%	80.20%	78.11%
General merchandise	66.78%	65.13%	70.56%	70.38%	75.89%	73.24%
Automobile	65.89%	63.78%	71.82%	74.35%	75.83%	74.56%
Hardware	60.63%	62.89%	64.78%	62.65%	65.86%	63.77%
Jewelry	62.95%	60.34%	69.15%	67.82%	73.33%	70.46%
Coating manufacturing	68.38%	65.76%	72.33%	71.65%	78.39%	76.22%
Western medicine	69.76%	66.32%	73.57%	73.22%	78.36%	75.42%
Medical equipment	64.56%	65.88%	72.94%	70.71%	73.64%	73.12%
Chemical products	65.27%	63.37%	70.22%	72.22%	73.20%	70.98%
Manufacturing products	62.66%	62.49%	69.17%	67.77%	72.41%	70.56%

The experimental results show that in a specific field, the average accuracy of ultra-short text classification method combined with the domain related data features and the recommendation algorithm is 74.71% , which is 9.19% and 3.81% higher than the traditional VSM and topic similarity method respectively , and the average recall is 72.64%, which is 8.36% and 2.39% higher than the two methods respectively.

From the experimental results above, the text classification method based on vector space model is less effective. The main reason is that the method only takes into account the word frequency statistics, generally speaking, for ultra-short text in specific field, it contains relatively few words, so that it is insufficient to embody its statistical advantages. The result of topic similarity method is better than the method based on vector space model. The reason is that it considers the semantic information of the word in the classified text. However, due to the scarcity of features in the ultra-short text in a specific field, the effect is unsatisfactory. The classification method combined the characteristics of the data and collaborative filtering ideas fully takes into account the feature information of the ultra-short text in the specific field, then achieves good results.

6 Conclusion

In this paper, ultra-short texts classification method combined characteristics of the data and collaborative filtering ideas in specific field is proposed. The experiment proves the effectiveness of the method. In the future work, we will further study how to get more internal information of data, then to further improve the accuracy of classification.

Acknowledgement

This research was supported by the State Key Laboratory of Digital Publishing Technology.

References

1. J. Bin. Micro-blog automatic classification method research and application. D. Harbin: Harbin Institute of Technology. (2012).
2. Y. Chao Chun. Short text categorization based on its own characteristics. D. HeFei University of Technology. (2016).
3. V.Bijalwan, V.Kumar, P.Kumari, et al. KNN based machine learning approach for text and document mining. J. *International Journal of Database Theory and Application.* **7,** 61-70, (2014).
4. V. Narayanan, I. Arora, A. Bhatia. Fast and accurate sentiment classification using an enhanced Naive Bayes model. M. *Intelligent Data Engineering and Automated Learning-IDEAL*. Springer Berlin Heidelberg, 194-201(2013).
5. B. Agarwal, N. Mittal. Text classification using machine learning methods-a survey. C. *Proceedings of the Second International Conference on Soft Computing for Problem Solving.* 701-709.(2014).
6. S. Maji, A.C.Berg, J. Malik. Efficient classification for additive kernel SVMs. J. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **35**, 66-77(2013).
7. A. Sun. Short text classification using very few words. C. *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval.* ACM, 1145-1146(2012).
8. D.BOLLEGALA, Y. MATSUO, M. ISHIZUKA. Measuring Semantic Similarity between Words Using Web Search Engines[EB/OL]. *http://ymatsuo.com/papers/jws-08danu.pdf,* (2016).
9. M. SAHAMI, T.D. HEILMAN. A web-based Kemel Function for Measuring the Similarity of Short Text Snippets. C. IEEE. International Conference on World Wide Web, 377-386(2016).
10. W. Sheng, F. Xinghua, C. Xianlin. Chinese short text classification using upper and lower relations. J. computer application, **30**, 603-606(2010).
11. F Yunjie, L Huailiang. Research on Chinese short text classification based on Wikipedia. J. modern library and information technology, **3**, 47-52(2012).
12. Z Hui. A Chinese short text classification algorithm based on Wikipedia. J. library and information work. **57**,120-124(2013).
13. X.H. Phan, L.M. Nguyen, S. Horiguchi. Learning to classify short and sparse text & web with hidden topics from large scale data collections. *In: Proceedings of the 17th International Conference on World Wide Web.* New York, USA:ACM, 91-100(2008).
14. Y Mengmeng, H Hao, C Luhong, M Ping, B Wu Jie. Short text classification based on LDA thematic model. J. computer engineering and design*,* **37**, 3371-3377(2012).
15. M Chenglong, Y Yonghong. Short text classification based on probabilistic semantic distribution. J. *automation journal*, **42** ,1711-1717(2016).