Applied Mechanics and Materials
ISSN: 1662-7482, Vols. 644-650, pp 2816-2820
doi:10.4028/www.scientific.net/AMM.644-650.2816

Submitted: 2014-07-20
Accepted: 2014-07-21
Online: 2014-09-22

A Combined Method for Chinese Micro-blogging Topic Tracking

Zhang Xiang, Shang Bo, Dong Lili, Zhu Yujie

College of Information and Control Engineering, Xi'an University of Architecture and Technology, Xi'an 710055, China

zhangxiang1001@126.com, yanliang12322@163.com, Donglilixjd@163.com, yujie210@126.com

Keywords: Chinese micro-blogging, topic tracking, LDA, Bagging

Abstract. To the problem of Chinese micro-blogging topic tracking, a method combined LDA model and Bagging of ensemble learning was proposed. The method firstly used the LDA hidden topic modeling, effectively solved the issue that the dataset's sparsity of the short text, then made the C4.5 decision tree as a weak classifier, through examples resampling to obtain multiple training set, compounding the training sets according to the voting rule, and ultimately getting the similarity of the micro-blogging topic. Experiments show that, compared with the model based on single vector model, classical TF-IDF and the tracking method of C.45Bagging similarity computing, this method have a better performance on precision, recall ratio and F1 value.

Introduction

As the Internet's development and Web2.0's emergency, micro-blogging has becoming an important tool for people to communicate with each other and make speech. Through micro-blogging, people can express their personal views, record their life knowledge and understand their relatives' recent news, at anytime, anywhere. Micro-blogging has the strong real-time character, which make the messages' propagation speed of the emergency happened in realistic society and hot news obviously faster than traditional media, such as television, newspaper and radio [1]. Therefore it is a hot issue in recent public opinion monitoring area that to find the hot topic in micro-blogging and to track them.

Topic tracking is one of the major research tasks of the topic discovery and tracking (Topic Detection and Tracking, TDT) [2], one example of information tracking, also one of the very important technologies of the new knowledge discovery, mainly solves the problem of 'what is the related information of the topic now and in the future', also means that in the context of determined in advance one or several topics and micro-blogging information related to these topics, according to a certain algorithm, tracking recognize the follow-up micro-blogging information belonged to a particular topic and publish comments, and to provide these micro-blogging information or comment's page link.

The traditional topic detection and tracking technology's research objects are mainly newswires, radio, television and online media which information has a long news report, the research data mainly obtains TREC meeting provided TDT corpus, since the scale of it is smaller. However, micro-blogging information has the character of short text, few feature word, huge scale and chatty language, traditional TDT technology cannot effectively be used to micro-blogging news.

So this paper adopts the combined potential Dirichlet Allocation model [3] (Latent Dirichlet Allocation, LDA) and the similarity calculation method of C4.5Bagging's micro-blogging hot topic tracking method, conducting the contrastive experiment on precision, recall and F1 value with traditional tracking method, and then analyze the results, make the conclusion that the new algorithm has a better tracking effect.

Text Process of Chinese micro-blogging topic tracking method

The detail process of Chinese micro-blogging topic tracking can be divided into the following steps: micro-blogging messages pretreatment, mainly contains corpus's choice, Chinese words' participle,

frequency count and so on; the feature representation and the extraction of the micro-blogging topic, means modeling the micro-blogging massage in a reasonable way, so that it's convenient to calculate the similarity between micro-blogging; the similarity calculation, is that use text classification to classify and evaluate the anticipated micro-blogging topic.

2.1.Data preparation and pretreatment.

Due to the current situation that there is no common data set in the micro-blogging data mining area, the Tencent micro-blogging's open API was chosen to be the information resource, randomly captured 181065 massages.

After a preliminary observation on these micro-blogging data, we find that there are many very smallish texts, primarily expression massages and interjection massages, still lots of Tencent game topic information. Therefore a basic clear management should be conducted, that is wipe off the massages whose length is less than 4 and topic is Tecent game, after the treatment there are 72162 massages left; then use the ICTCLAS word segmentation system of Chinese academy of science, conduct the segmentation treatment of the micro-blogging information; finally make the stop word processing.

2.2.Micro-blogging's representation model based on LDA.

After the pretreatment is finished, it's time to show the result in a computable way to a computer, this process is called the micro-blogging's representation model. This paper use LDA (Latent Dirichlet Allocation) to be the micro-blogging topic's representation model. LDA is a three generation dendritic Bayesian probability model, consisted by a document layer, a subject layer and a word layer, as is shown in figure 1. The LDA model bases on the assumption that 1. There are K separated theme in one document set; 2. Every theme is the multinomial distribution over words; 3. Document is mixed by K themes; 4. Every document is the multinomial distribution over k themes; 5. Every document's distribution probability is generated by Dirichlet distribution.

In the Bayes probability model of generating three-tree, is a K-dimensional vector, which represents the polynomial distribution of the document set over themes, assume that $\theta = [\theta_1, \theta_2, ..., \theta_k]$,then $\theta_1 + \theta_2 + ...\theta_k = 1$ and $0 \le \theta_k \le 1$, $1 \le k \le K$. α is the K-dimensional Dirichlet super parameters set corresponding to the document set, where $\alpha = [\alpha_1, \alpha_2, ..., \alpha_k]$. LDA generated model can be represented using a Bayesian network diagram.

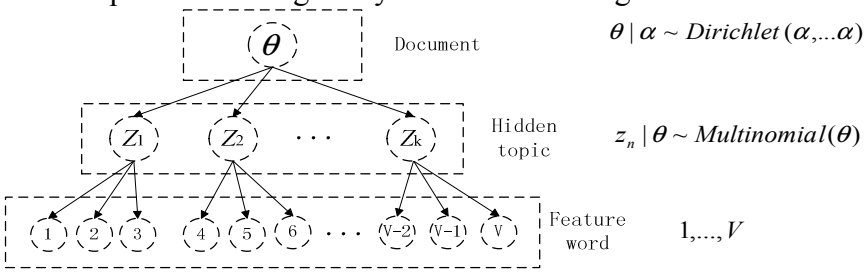

Fig.1 LDA Bayesian probability generation model

To a document set, document-level parameters α and β are sampled once; to every document set, the document-level parameter θ is sampled once too; so as to the word-level parameters z and w , the every word in the document. Since the LDA model's generation process is as follows:

The first step is extracting N words, N obeys Poisson distribution;

The second step is extracting θ (the vectors that topic occurrence's possibility), obeys Dirichlet distribution;

The third step is that, to every w_n in N words:

1. Generating the theme z_n , z_n abides by multinomial distribution.
2. Generating z_n from $p(w_n \mid z_n, \beta)$, which abides the theme z_n condition.

Where β is a $k \times V$ matrix, representing as $\beta = p(w^j = 1 \mid z^i = 1)$ recording a certain word's generation probability at a certain topic, is a compound variable. More introductions about LDA model please look at the reference [3].

2.3 Micro-blogging topic tracking based on C4.5Bagging

After using the LDA model to explain the micro-blogging, computer can conduct calculation work on the micro-blogging topic. This paper uses the C4.5Bagging algorithm to calculate the similarity over unknown micro-blogging and the certain micro-blogging topic. If the similarity belongs to the theme, then we can track the micro-blogging topic.

2.3.1 Bagging theory

Generalization capability has always been a fundamental problem in machine learning. Since ensemble learning can effectively improve the generalization ability, becoming a hot point in machine learning [4]. The current ensemble learning algorithm can be roughly divided into two types, namely is no strong dependency between the individual learner, can parallel generate algorithm, another type is that there is a strong dependence between individuals, must generate serial algorithm. The former's representative algorithm is Bagging algorithm, while the latter is Boosting algorithm. Since the Boosting algorithm dealing with some practical issues will lead to over-fitting problems, may obtains the worse classified results than single weak classifier, this paper's integrated learning choose the Bagging algorithm.

The basic idea of Bagging algorithm is:

Bagging the basic idea is: Given a weak classifier and a training set, let the weak classifier train T rounds, each round of training set randomly fetch n training samples from the initial training set, after every training gains a prediction function h_i, at the same training, there are T prediction functions h_1, h_2……, h_T. Using this prediction function list to forecast the sample set, then get the final prediction result h^* according to the majority voting rules.

2.3.2 C4.5 integrated learning based on Bagging

Breiman pointed out that in Bagging, stability is one of the key factors to improve the prediction accuracy, the so called 'unstable' means: a small change in the data set can cause significant changes in the classification results. According to Beriman's research, the decision trees, neural networks are unstable, Bagging algorithm can rise the prediction accuracy to these unstable learning algorithm, while the effect on the stable is less obvious, sometimes even decline the accuracy [5]. This paper chooses the selection tree as the integrated learning's weak classifiers.

In data mining techniques, decision tree is a simple and efficient method compared with neural networks and Bayesian methods, decision trees do not spend a lot of time and thousands of times to iterate the training model, do not need external information except the information about the training data, showing good classification accuracy. The basic algorithm of decision tree induction algorithm is greedy algorithm, it use top-down way to construct the decision tree. The main representative algorithms are ID3, C4.5 algorithm and so on. This paper selects the C4.5 algorithm to establish decision tree. C4.5 algorithm adopts the way that from top to bottom to handle the problem, which is developed on the basis of ID3 algorithm, and has many significant improvement compared with ID3 algorithm, including that it can deal with the example with numerical attributes and the some missing data, and it uses gain ratio to choose the attribute.

Using Bagging method to integrate the C4.5 algorithm, its main idea is as follows:

Set $S = \{x_1, y_1\}, \{x_2, y_2\}, \ldots \{x_m, y_n\}$, as the training set, a total number of m documents are divided into n classes. In which x_i is sample, y_i is the text category. Using the repeatable sampling techniques obtains T new training sets $\{S_j | j = 1,2,\ldots T\}$ whose scale is closed to the original training set. Due to the repeatable sampling, some examples in the original training set may appear several times in the new training set, while some others may never appears. After that, using C4.5 algorithm to make every new training set to learn, through the voting rules, integrating the learning results. The algorithm is described as follows:

Input: Training Set S, Machine Learning Algorithm C4.5, Integer T (number of iteration)
For i=1 to T {

$S_i = resampling(S)$ //using the resampling techniques obtains new training sets S_i

$H_i = C4.5(S_i)$ //applying C4.5 to train the training set S_i

}

$H^*(x) = \arg\max_{y \in Y} \sum_{i=1}^{T} I(H_i(x) = y)$ // Ensemble learning result through voting rule

Output: classifier H^* //return the result

2.3.3 Performance index

There are similar points between TDT research and IR research, this paper takes over the IR area's evaluation criterion to evaluate the precision/recall and the F value. The precision is calculated by the way as is shown in formula (1), the recall is shown in formula (2):

$$precision = \frac{P_T}{P_T + P_F} \tag{1}$$

$$recall = \frac{P_T}{P_T + N_T} \tag{2}$$

In which P_T is the number of the document that is correctly assign to this class, P_F is the number of the document that is mistakenly assign to this class, N_T is the number of document that should be send to the class but not be send. F value composite the precision and recall radio, which can general evaluate the classifier, the common F1 measurement index (referred as F1) is shown as formula (3):

$$F_1 = \frac{precision * recall * 2}{precision + recall} \tag{3}$$

Experimental result and analysis

The experimental environment is as follows: CPU: 4 * Intel Core I7-2600@3.40G; Memory: 4G; HDD: 1T 7200 RPM IDE; Operating System: Windows7; But to the Chinese micro-blogging, there is no standard Chinese corpus. This paper divided the data collected from the Tencent micro-blogging open API into 5 topics. Every topic selects 1000 news, 700 news in each topic will be designated as a training sample, another 300 News designated as test samples.

Currently in the research of topic tracking, the most widely used benchmark system is the tracking method based on single vector representation model, the most classical combination of this method is single vector representation model, classical TF*IDF and the cosine similarity calculation. The experimental task of this paper is to make a comparison between the classical combination and C4.5 Bagging method over accuracy, recall and the F1 value. During the experiment: for the parameter estimation of LDA model, this paper adopts Gibbs [6] sampling method, according to reference [7], the number of hidden topic is 200, the initial super parameter α is 0.25, and β is 0.1. In C4.5Bagging method, the value of T = 20. The experimental result is as shown in table 1.

Table1 The evaluation results of three algorithms' topic tracking

algorithm	precision	recall	F1
single vector representation model	0.798	0.726	0.760
TF* IDF and C.45Bagging	0.828	0.794	0.811
LAD and C.45Bagging	0.862	0.819	0.835

From the test results, whether it is accuracy, recall or F1 values, LDA and C4.5Bagging tracking classification used in this paper have the best effect, C4.5Bagging algorithm's final classification decision is given to a number of weak classifiers decisions, and get to overcome the problem of instability in a single classifier, so the Bagging method can improve the prediction accuracy of unstable learning algorithm. While due to LDA topic model has simple parameters, it is not generate over-fitting, therefore, it can be a good representation of micro-blogging topic's theme.

Conclusion

This paper studies the hot topics' tracking methods of Chinese micro-blogging, using the hidden theme LDA modeling method, effectively solving the sparse problem of the short text dataset, then using Bagging algorithm as a classification algorithm designed a micro-blogging topic tracking methods, improved the tracking efficiency in some way. Next work mainly in the following two aspects: first, to solve the drift problem in micro-blogging topic; Second, the iteration number of Bagging algorithm determined mainly by experience, therefore it has a less adaptive capacity, which will be a further issues to be addressed.

Acknowledgment

Natural Science Foundation of Shaanxi Provincial under Grant No. 2012JM8042; Natural Science Foundation of Shaanxi Provincial Department of Education under Grant No. 12JK0940; Xian Project to Promote Technology Transfer under Grant No. CXY1348-(1) ; Science and Technology Plan Projects of Yulin city under Grant No. 12-2-07.

References

[1] Http://baike.baidu.com~iew/1259292.

[2] Zhang Xiaoyan, Wang Ting, Liang Xiaobo. Use of LDA Model in Topic Tracking [J]. Computer Science, 2011, 38(10):136-139

[3] Blei D M, Ng A Y. Latent Dirichlet Allocation [J]. The Journal of Machine Learning Research, 2003, 3:993－1022.

[4] Zhang Xiang, Zhou Mingquan, Geng Guohua. Research on Improvement of Bagging Chinese Text Categorization Classifier [J]. Journal of Chinese Computer Systems, 2010, 31(2): 281-284

[5] Shen Xuehua, Zhou Zhihua, Wu Jianxin, Chen Zhaoqian. Survey of Boosting and Bagging[J]. Computer Engineering and Applications , 2000, 36(12): 31-32

[6]Thomas L. Griffiths, Mark Steyvers. Finding scientific topics[J].Proceedings of the National Academy of Sciences of the United States of America,2004,101(suppl 1):5228-5235.

[7] Lu Rong, Xiang Liang, Liu Mingrong, Yang Qing. Discovering News Topics from Micro-blogs Based on Hidden Topics Analysis and Text Clustering [J]. Pattern Recognition and Artificial Intelligence, 2012, 03:382-387.

[8] Zhang Xiang, Zhou Ming-quan, Geng Guo-hua.C4.5 Bagging algorithm for Chinese text categorization. Computer Engineering and Applications,2009,45(26):135-137.

Machine Tool Technology, Mechatronics and Information Engineering
10.4028/www.scientific.net/AMM.644-650

A Combined Method for Chinese Micro-Blogging Topic Tracking
10.4028/www.scientific.net/AMM.644-650.2816