Towards Effective Web Page Classification

Min Gu, Feng Zhu, Qing Guo, Yanhui Gu*, Junsheng Zhou, Weiguang Qu
School of Computer Science and Technology
Nanjing Normal University
gu@njnu.edu.cn

***Abstract*—In order to manage and organize information on the web, we propose a novel web page classification strategy integrating topic model and SVM. We use topic model to harness the implicit information on web pages for feature extraction. Accuracy of the strategy is 84.15%, 2.23% superior to the traditional classification strategy based on CHI.**

I. Introduction

Due to the rapid development of web documents, web page classification has become one of the key techniques for managing and organizing those information on the Web. It can help Information Retrieval and Digital Monitoring. The traditional web page classification strategies are based on plain text classification methods. Most previous works use BOW(Bag of Words) model to represent features. However, it ignores the synonyms and may cause curse of dimensionality. Also, the traditional feature extraction may lose some features with low frequency, which may lead to bad performance.

In this paper, we propose a web page classification strategy integrating topic model. We utilize the latent information in the web pages for feature extraction. In topic model, terms are clustered to some latent topics. We apply the Latent Dirichlet Allocation(LDA) algorithm to generate a probabilistic topic model from the web page collection. It will reduce the number of feature dimensions and map the semantically related terms into same dimension. It can also help the low-frequent terms perform well in the classification. We utilize the Support Vector Machines for classification. We compare the classification strategy based on CHI with the classification strategy integrating topic model. Results of the experiments show that the strategy is efficient on the crowd undertaking corpus and other corpus.

II. Preliminary

A. Topic model

Topic model is the modeling method for implicit topics in text. The most classic topic model is LDA[1]. It is a three-tier Bayesian probability model including words, topics and document structures. It is an unsupervised probabilistic model and is widely used to discover latent semantic structure of a document collection. Some works have attempted to use LDA in web page classification. In recent years, LDA has been used in unsupervised text classification, such as [2] and [3].

B. Web page classification

The classification algorithms used in web page classification include Naive Bayes, SVM, and Neural Network. In feature extraction, latent semantic analysis has been used for reducing the number of feature dimensions[4]. It has an improvement of 23.31% comparing to BOW model.

III. Web page classification strategy integrating topic model

The framework of our strategy is shown in Fig. 1. First, we extract text from web pages and segment the text. Then we extract features from latent topics by using multi-LDA. Multi-LDA is the collection of LDAs for different categories. We apply it to extract topics and words from different categories respectively. So the interaction between categories will be reduced. Finally, we train the model by using SVM algorithm and predict new examples.

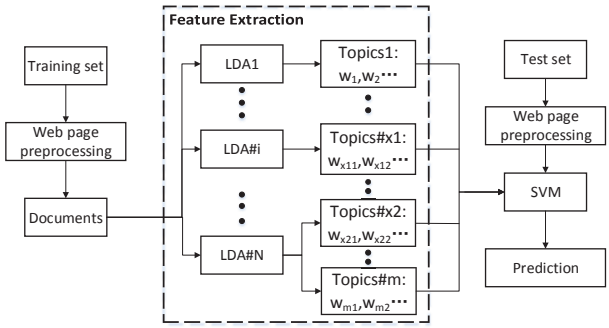

Fig. 1. The framework of classification strategy

A. Web page preprocessing

We utilize the extraction method based on DOM Tree(Document Object Model Tree). We can obtain the attitude information and text of the tags through DOM Tree. There are many noise in web pages, such as advertisements. We use Jsoup Parser to clean the pages. Then, we can get the plain text and use Ansj toolkit to segment text.

B. Feature extraction

Feature extraction aims to find out features which have strong ability to distinguish categories in the training set. Traditional feature extraction method including CHI, IG, and MI are based on statistics which ignore the semantic characteristics of terms. In this paper, we utilize the latent information in web pages to obtain the features through Multi-LDA, the collection of LDAs for categories.

Given the hyperparameter $\overrightarrow{\alpha}$ and matrix parameter $\overrightarrow{\beta}$, we can calculate the joint distribution of a topic mixture $\overrightarrow{\theta}$. In

the learning, we use Gibbs sampling to estimate approximate posterior inference in LDA. The generative process of LDA is shown in Algorithm1.

We apply multi-LDA for different categories. Using multi-LDA, we can obtain the latent topics and words belong to the topics based on the labeled data in web page collection. We use words as features for classification. Then, we measure the weight of features using TF-IDF.

Algorithm 1 generative process of LDA

for all topics $k \in [1, K]$ **do**
  sample mixture components $\overrightarrow{\varphi_k} \sim Dir(\overrightarrow{\beta})$
end for
for all documents $m \in [1, M]$ **do**
  sample mixture proportion $\overrightarrow{\theta_m} \sim Dir(\overrightarrow{\alpha})$
  document lenghth $N_m \sim Poiss(\xi)$
  for all words $n \in [1, N_m] in\, document\, m$ **do**
    sample topic index $z_{m,n} \sim Mult(\overrightarrow{\theta_m})$
    sample term for word $w_{m,n} \sim Mult(\overrightarrow{\varphi}_{m,n})$
  end for
end for

C. Web page classification

After web page preprocessing and feature extraction, we obtain the original training set. We choose SVM as the classifier. After training, we can get the model for classification. In the paper, we use libsvm toolkit. Each test web page can be assigned to the category getting the hightest number of votes based on the model.

IV. Experiments

A. Datasets

We use Chinese Sogou web page collection and the crowd undertaking web page collection. The Sogou corpus has 20,000 web pages divided in 8 categories (16,000 for training, 4000 for test). The crowd undertaking corpus is collected from the Internet. It contains 2,264 pages covering 4 categories (1816 pages for training, 448 pages for test).

B. Evaluation metrics

To compare the result of the strategy based on CHI with the strategy integrating topic model, we choose the traditional evaluation standards: Accuracy, Precision, Recall, and F-measure.

$$Accuracy = \frac{\#correctly\ classified\ test\ pages}{\#all\ test\ pages} \quad (1)$$

$$Precision = \frac{\#correct\ positive\ predictions}{\#positive\ predictions} \quad (2)$$

$$Recall = \frac{\#correct\ positive\ predictions}{\#positive\ data} \quad (3)$$

$$F - measure = \frac{2 * Precision * Recall}{Precision + Recall} \quad (4)$$

C. Evaluation of Performance

We compare the web page classification strategy using CHI with the strategy using multi-LDA on the crowd undertaking corpus. The result is shown in Tabel. I.

TABLE I
The Examples Of Templates

#Features	Strategy based on CHI(Accuracy)	Strategy based on Topic Model(Accuracy)
1000	84.15%	76.56%
2000	82.81%	82.14%
3000	82.37%	82.59%
4000	82.81%	83.04%
5000	81.92%	84.15%

From Fig.2, the results using multi-LDA yield a higher precision compared to applying CHI in feature extraction. Multi-LDA extracts features considering semantic characteristics.

We also observe the relation between the number of topics and the classification performance on the Sogou corpus in Fig. 2.

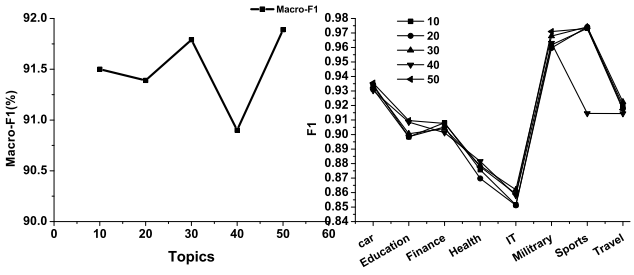

Fig. 2. Results of Sogou corpus

We set the number of features equal to 4000. We find that the classification performance is better with increase of topic numbers. However, the first curve in Fig.3 is flat while the cost of time is increasing.

V. Conclusion

We propose the web page classification strategy integrating topic model. With the latent information obtained from multi-LDA, the relation between features and documents becomes closer. In the future, we will do more research on unsupervised web page classification.

Acknowledgment

We would like to thank the anonymous reviewers for their insightful comments. This work is supported by Chinese National Fund of Natural Science under Grant 61272221, 61472191,the Natural Science Research of Jiangsu Higher Education Institutions of China under Grant 14KJB520022.

References

[1] D. M Blei, A. Y. Ng and M. I. Jordan. *Latent dirichlet allocation*, Journal of Machine Learning Research, 2003, 3: 993-1022.

[2] R. Fu, B. Qin and T. Liu. *Open-categorical text classification based on multi-LDA models*, Soft Computing, 2014, 19(1): 29-38.

[3] S. Hingmire and S. Chakraborti. *Sprinkling Topics for Weakly Supervised Text Classification*, Processings of the Association for Computational Linguistics, 2014: 55-60.

[4] S. Makoto, S. Akio and M. Tohru. *Hierarchical Web Page Classification Based on a Topic Model and Neighboring Pages Integration*, Eprint Arxiv, 2010, 64(5): 1973-1979.