IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 12, December 2015

Fine Grained Model Based Relevance Feature Discovery for Text Mining

S. H. Jadhav[1], P. B. Koli[2]

PG Student, Late GNSCOE, Nasik, Maharashtra, India[1]

Professor, Computer Engineering Department, Late GNSCOE, Nasik, Maharashtra[2]

Abstract: For describing user preferences because of large scale terms and data patterns, it is a big challenge to guarantee the quality of discovered relevance features in text documents. Term-based approaches, most existing popular text mining and classification methods have adopted. However, they have all suffered from the problems of polysemy and synonymy. Yet, how to effectively use large scale patterns remains a hard problem in text mining, over the years, there has been often held the hypothesis that pattern-based methods should perform better than term-based ones in describing user preferences; This paper presents an innovative model for relevance feature discovery, to make a breakthrough in this challenging issue. As higher level features and deploys them over low-level features (terms), it discovers both positive and negative patterns in text documents. On their specificity and their distributions in patterns, it also classifies terms into categories and updates term weights.

Keywords: Text mining; text feature extraction; text classification;

1. INTRODUCTION

Text classification is a popular and important text mining task. Many document collections are multi-class and some are multi-label. Both multi-class and multi-label data collections can be dealt with by using binary classifications. For finding and classifying low-level terms based on both their appearances in the higher-level features (patterns) and their specificity in a training set, this paper proposes an innovative technique. To select irrelevant documents (so-called offenders) that are closed, it also introduces a method. The advantages of the proposed model as compared with other methods:

- Effective use of both the feedback (i.e. relevant and irrelevant feedback) to find useful features;
- Integration of term and pattern features together instead of using them in two different stages.

The first RFD model uses two empirical parameters to set the boundary between the categories. It achieves the expected performance, but the manually testing of a large number of different values of parameters is not comfortable. For example as in [20], word "LIB" may be more frequently used than word "JDK"; but "JDK" is more specific than "LIB" for describing "Java Programming Languages"; and "LIB" is more general than "JDK" because "LIB" is also frequently used in other programming languages like C or C++. Therefore, we recommend the consideration of both terms' distributions and specificities for relevance feature discovery. The research presents a method to find and classify low-level features based on both their appearances in the higher-level patterns and their specificity. It also introduces a method to select irrelevant documents for weighting features.

2. RELATED WORK

Feature selection and feature extraction are the most important steps in classification systems. Feature selection is commonly used to reduce dimensionality of datasets with tens or hundreds of thousands of features which would be impossible to process further. One of the problems in which feature selection is essential is text categorization. A major problem of text categorization is the high dimensionality of the feature space; therefore, feature selection is the most important step in text categorization. At present there are many methods to deal with text feature selection. Ant colony optimization algorithm is inspired by observation on real ants in their search for the shortest paths to food sources. Proposed algorithm is easily implemented and because of use of a simple classifier in that, its computational complexity is very low. The performance of proposed algorithm is compared to the performance of genetic algorithm, information gain and CHI on the task of feature selection in Reuters-21578 dataset. Simulation results on Reuters-21578 dataset show the superiority of the proposed algorithm.

Term-based approaches can extract many features in text documents, but most include noise. Many popular text-mining strategies have been adapted to reduce noisy information from extracted features; however, text-mining techniques suffer from low frequency. The key issue is how to discover relevance features in text documents to fulfill user information needs. To address this issue, we propose a new method to extract specific features from user relevance feedback. The proposed approach includes two stages. The first stage extracts topics (or patterns) from text documents to focus on interesting topics. In the second stage, topics are deployed to lower level terms to address the low-frequency problem and find specific terms. The specific terms are determined based on their appearances in relevance feedback and their distribution in topics or high-level patterns.

Relevance Feedback (RF) has been proven very effective

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 12, December 2015

for improving retrieval accuracy. Adaptive information altering (AIF) technology has benefited from the improvements achieved in all the tasks involved over the last decades. A difficult problem in AIF has been how to update the system with new feedback efficiently and effectively. In current feedback methods, the updating processes focus on up- dating system parameters. In this paper, we developed a new approach, the Adaptive Relevance Features Discovery (ARFD). It automatically updates the system's knowledge based on a sliding window over positive and negative feedback to solve a non-monotonic problem efficiently. Some of the new training documents will be selected using the knowledge that the system currently obtained. Then, specific features will be extracted from selected training documents. Different methods have been used to merge and revise the weights of features in a vector space.

3. PROPOSED SYSTEM

The new model is designed for Relevance Features Discovery (RFD), a pattern mining based approach, which uses negative relevance feedback to improve the quality of extracted features from positive feedback. Learning algorithms are also proposed to implement this approach on Reuters Corpus Volume 1 and TREC topics. Experiments show that the proposed approach can work efficiently and achieves the encouragement performance.
We present a document classification system that employs lazy learning from labeled phrases, and argue that the system can be highly effective whenever the following property holds: most of information on document labels is captured in phrases. We call this property *near sufficiency*.

Our research contribution is twofold: (a) we quantify the near sufficiency property using the Information Bottleneck principle and show that it is easy to check on a given dataset; (b) we reveal that in all practical cases from small-scale to very large-scale manual labeling of phrases is feasible: the natural language constrains the number of common phrases composed of a vocabulary to grow *linearly* with the size of the vocabulary. Both these contributions provide firm foundation to applicability of the phrase-based classification (PBC) framework to a variety of large-scale tasks. We deployed the PBC system on the task of *job title classification*, as a part of LinkedIn's data standardization effort. The system significantly outperforms its predecessor both in terms of precision and coverage. It is currently being used in LinkedIn's ad targeting product, and more applications are being developed. We argue that PBC excels in high explain ability of the classification results, as well as in low development and low maintenance costs. We benchmark PBC against existing high-precision document classification.

In this survey we review work in machine learning on methods for handling data sets containing large amounts of irrelevant information We focus on two key issues the problem of selecting relevant features and the problem of selecting relevant examples We describe the advances that have been made on these topics in both empirical and theoretical work in machine learning and we present a general framework that we use to compare different methods We close with some challenges for future work in this area.

In the proposed work, we will use fine-grained topic modeling approach to jointly identify opinion features, including non-noun features, infrequent features, as well asimplicit feature. In addition, neutral opinions will be considered; currently only positive and negative opinions are considered.

WFeature ranks documents by using a set of extracted features. Feature Weighting gives the degree of information represented by the feature occurrences in a document and reflects the relevance of the feature. FCluster classifies the ranked features into multiple classes. Rather than use all irrelevant documents, some offenders (i.e., top-K ranked irrelevant documents) are selected.

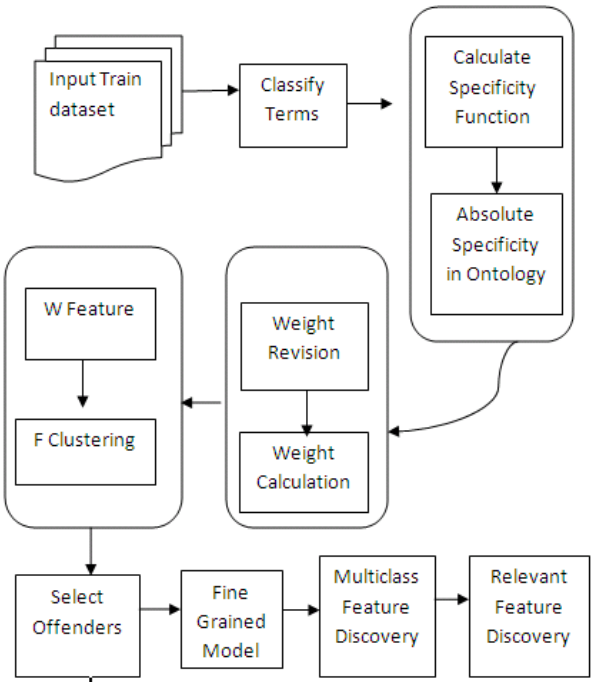

Fig. 1: Overview of Proposed System

4. CONCLUSION

The research proposes an alternative approach for relevance feature discovery in text documents. It presents a method to find and classify low-level features based on both their appearances in the higher-level patterns and their specificity. It also introduces a method to select irrelevant documents for weighting features. In this paper, we continued to develop the RFD model and experimentally prove that the proposed specificity function is reasonable and the term classification can be effectively approximated by a feature clustering method.
The first RFD model uses two empirical parameters to set the boundary between the categories. It achieves the expected performance, but it requires the manually testing of a large number of different values of parameters. The new model uses a feature clustering technique to automatically group terms into the three categories.

DOI 10.17148/IJARCCE.2015.412131

Compared with the first model, the new model is much more efficient and achieved the satisfactory performance as well. This paper also includes a set of experiments on RCV1 (TREC topics), Reuters-21578 and LCSH ontology. These experiments illustrate that the proposed model achieves the best performance for comparing with term-based baseline models and pattern-based baseline models. The clustering method is effective and the proposed model is robust, the results show that the proposed specificity function is adequate.

This paper demonstrates that the proposed model was thoroughly tested and the results prove that the proposed model is statistically significant. The paper also proves that the use of irrelevance feedback is important for improving the performance of relevance feature discovery models. It provides a promising methodology for developing effective text mining models for relevance feature discovery based on both positive and negative feedback.

ACKNOWLEDGMENT

I thank to my project guide **Prof. P. B. Koli** and ME Coordinator **Prof. J. V. Shinde** , Our Head of Computer Department **Prof. N. R. Wankhade** and our faculty member of computer Engineering, Late GNS college of Engineering Nasik. Without their importance this paper could not be completed.

REFERENCES

[1] M. Aghdam, N. Ghasem-Aghaee, and M. Basiri, "Text feature selection using ant colony optimization," in Expert Syst. Appl., vol. 36, pp. 6843–6853, 2009.
[2] A. Algarni and Y. Li, "Mining specific features for acquiring user information needs," in Proc. Pacific Asia Knowl. Discovery Data Mining, 2013, pp. 532–543.
[3] A. Algarni, Y. Li, and Y. Xu, "Selected new training documents to update user profile," in Proc. Int. Conf. Inf. Knowl. Manage., 2010, pp. 799–808.
[4] N. Azam and J. Yao, "Comparison of term frequency and document frequency based feature selection metrics in text categorization," Expert Syst. Appl., vol. 39, no. 5, pp. 4760–4768, 2012.
[5] R. Bekkerman and M. Gavish, "High-precision phrase-based document classification on a modern scale," in Proc. 11th ACM SIGKDD Knowl. Discovery Data Mining, 2011, pp. 231–239.
[6] A. Blum and P. Langley, "Selection of relevant features and examples in machine learning," Artif. Intell., vol. 97, nos. 1/2, pp. 245– 271, 1997.
[7] C. Buckley, G. Salton, and J. Allan, "The effect of adding relevance information in a relevance feedback environment," in Proc. Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 1994, pp. 292–300.
[8] G. Cao, J.-Y. Nie, J. Gao, and S. Robertson, "Selecting good expansion terms for pseudo-relevance feedback," in Proc. Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2008, pp. 243–250.
[9] G. Chandrashekar and F. Sahin, "Asurvey on feature selection methods," in Comput. Electr. Eng., vol. 40, pp. 16–28, 2014.
[10] B. Croft, D. Metzler, and T. Strohman, Search Engines: Information Retrieval in Practice. Reading, MA, USA: Addison-Wesley, 2009.
[11] S. Zhu, X. Ji, W. Xu, and Y. Gong, "Multi-labelled classification using maximum entropy method," in Proc. Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2005, pp. 1041–1048.
[12] N. Zhong, Y. Li, and S.-T. Wu, "Effective pattern discovery for text mining," in IEEE Trans. Knowl. Data Eng., vol. 24, no. 1, pp. 30–44, Jan. 2012.
[13] Z. Zhao, L. Wang, H. Liu, and J. Ye, "On similarity preserving feature selection," in IEEE Trans. Knowl. Data Eng., vol. 25, no. 3, pp. 619–632, Mar. 2013
[14] S. Shehata, F. Karray, and M. Kamel, "A concept-based model for enhancing text categorization," in Proc. ACM SIGKDD Knowl. Discovery Data Mining, 2007, pp. 629–637.
[15] I. Soboroff and S. Robertson, "Building a filtering test collection for TREC 2002," in Proc. Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2003, pp. 243–250.
[16] F. Song and W. B. Croft, "A general language model for information retrieval," in Proc. ACM Conf. Inf. Knowl. Manage., 1999, pp. 316–321.
[17] Q. Song, J. Ni, and G. Wang, "A fast clustering-based feature subset selection algorithm for high-dimensional data," in IEEE Trans. Knowl. Data Eng., vol. 25, no. 1, pp. 1–14, Jan. 2013.
[18] Y. Li, D. F. Hus, and S. M. Chung, "Combination of multiple feature selection methods for text categorization by using combinational fusion analysis and rank-score characteristic," Int. J. Artif. Intell. Tools, vol. 22, no. 2, p. 1350001, 2013.
[19] Y. Li, X. Zhou, P. Bruza, Y. Xu, and R. Y. Lau, "A two-stage text mining model for information filtering," in Proc. 17th ACM Conf. Inf. Knowl. Manage., 2008, pp. 1023–1032.
[20] Y. Li, A. Algarni, and N. Zhong, "Mining positive and negative patterns for relevance feature discovery," in Proc. ACM SIGKDD Knowl. Discovery Data Mining , 2010, pp. 753–762
[21] Y. Li, X. Zhou, P. Bruza, Y. Xu, and R. Y. Lau, "Two-stage decision model for information filtering," Decision Support Syst., vol. 52, no. 3, pp. 706–716, 2012.

BIOGRAPHIES

Mrs. Jadhav Suvarna H. received the B.E in Computer Engineering from SRES COE, KOPARGAON, Maharashtra. Currently perusing M.E from Late GNSCOE, Nasik, Maharashtra .

Prof. P. B. Koli M. Tech. Computer Science, BE (COMP). Professor at Late GNSCOE, Nasik, Maharashtra.

DOI 10.17148/IJARCCE.2015.412131