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

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 9, September 2015

A Survey on Concept based Pattern Discovery for Text Mining

Muthuvalli.A.R[1], Manikandan.M[2]

PG Scholar, Dept of Computer Science and Engineering, Adhiyamaan College of Engineering, Hosur, India[1]

Assistant Professor, Dept of Computer Science and Engineering, Adhiyamaan College of Engineering, Hosur, India[2]

Abstract: Text Mining is the technique that helps users find useful information from a large amount of text documents on the web or database.Most popular text mining and classification methods have adopted term-based approaches. The term based approaches and the pattern-based method describing user preferences; yet, how to effectively use large scale patterns remains a hard problem in text mining. This paper discovered two challenging issues like low-support-problem and misinterpretation problem, to make a breakthrough in these challenging issues, presents an innovative model that is concept-based model. A concept-based mining model that relies techniques on the analysis of both the sentence based and document based concept analysis to overcome the issues of low-support-problem and misinterpretation problem and to improve the effectiveness of using and updating discovered patterns for finding relevant and irrelevant documents.

Keywords: Text mining, Text feature extraction, Text classification, Pattern mining.

I. INTRODUCTION

Relevance Feature Discovery (RFD) is to find the useful features available in text documents including both relevant and irrelevant documents.Text mining is the discovery of interesting knowledge in text documents. It is a challenging issue to find accurate knowledge (or features) in text documents to help users to find what they want. It discovered two challenging issues in using pattern mining techniques for finding relevance features in both relevant and irrelevant documents. The first issue is the low-support problem. If the minimum support is decreased, a lot of noisy patterns can be discovered. The second issue is the misinterpretation problem. The measures (e.g., “support” and “confidence”) used in pattern mining turn out to be not suitable in using patterns for solving problems. Concept-based model (CBM) is used to solving these types of issues. It discovers concepts by using natural language processing (NLP) techniques. It is a verb-argument structures to find concepts in sentences.To make a breakthrough for the two challenging issues, presents an innovative model it provides specificity function and used two empirical parameters to group terms into three categories: “positive specific terms”, “general terms”, and “negative specific terms”based on the RFD model.

The remainder of the paper is organized as follows. Section 2 shows the overview of the related work. In section 3 shows the methodologies of the survey. And Section 4 shows the survey conclusion.

II. RELATED WORK

I. Relevance Feature Discovery for Text Mining

In this paper FClustering and WFeature algorithms are used.Algorithm FClustering describes the process of feature clustering,where DP^+ is the set of discovered patterns of D^+ and DP^- is the set of discovered patterns of D^-.Algorithm WFeature is used to calculate term weights after terms are classified using Algorithm FClustering.

II.Effective Pattern Discovery for Text Mining

In this paper D-Pattern Mining Algorithm are used. It describes the training process of finding the set of d-patterns. The main focus of this paper is the deploying process, which consists of the d-pattern discovery and term support evaluation. It discovered all patterns in a positive document are composed into a d-pattern.

III.Mining Positive and Negative Patterns for RelevanceFeature Discovery

In this paper two algorithms are used that is Mining and Revision algorithms.The algorithm calls twice one for positive documents and one for negative documents. The process of the revision firstly finds features in the positive documents in the training set, including higher level positive patterns and low-level terms. It then selects top-K negative samples in the training set according to the positive features. It also discovers negative patterns and terms from selected negative documents using the same pattern mining technique that we used for the feature discovery in positive documents. In addition, the process revises the initial features and obtains a revised weight function. The former finds higher level positive features, selects top-K negative samples, discovers higher level negative features, and composes the set of terms.

IV.An Efficient Concept-Based Mining Model for Enhancing Text Clustering

In this paper Concept-Based Analysis Algorithm are used. The concept-based analysis algorithm describes the concepts in the documents. Each concept in the verb

DOI 10.17148/IJARCCE.2015.4993

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

argument structures, which represents the semantic structures of the sentence, is processed sequentially. Each concept in the current document is matched with the other concepts in the previously processed documents.

V. A Topic based Document Relevance Ranking Model

In this paper Pattern Enhanced Topic Model (PETM) are used. PETM determine document relevance based on topics distribution and maximum matched patterns. LDA (Latent Dirichlet Allocation) is one of the most popular probabilistic text modelling techniques.It can discover the hidden topics in collections of documents with the appearing words. PETM pattern mining is used to discover semantically meaningful and efficient patterns torepresent topics and documents are implemented in two steps. Firstly, construct a new transactional dataset from the LDA outcomes of the document collection; secondly, generate pattern based representations from the transactional dataset to represent user needs.

VI.On Similarity Preserving Feature Selection

In this paper SPFS-NES (Similarity Preserving Feature Selection-Nesterov's method) are used.Feature selection isto choose a subset of the original features according to a selection criterion. It selects a small set of the original features. The original features, feature selection improves the interpretability of learning models and it is a learning process, and use the objective function of the learning model to guide searching for relevant features.

VII.High Dimensional Data Clustering using Fast Cluster Based Feature Selection

In this paper Fast clustering based feature Selection algorithm (FAST) are used. Based on the MST method the FAST algorithm works in two steps. In the first step, features are divided into clusters by using graph-theoretic clustering methods. In the second step, the most representative feature that is strongly related to target classes is selected from each cluster to form the final subset of features.

VIII. A Two-stage Information Filtering Based onRough Decision Rule and Pattern Mining

In this paper two stage information filtering are used. In first filtering stage issupported by a novel rough analysis model which efficiently removes a large number of irrelevant documents, thereby addressing the overload problem. In second filtering stage is empowered by a semantically rich pattern taxonomy mining model which effectively fetches incoming documents according to the specific information needs of a user, thereby addressing the mismatch problem.

III. METHODOLOGY

In this paper we introduce the RFD (Relevance Feature Discovery) model for concept based method. It describes the relevant features in relation to three groups: positive specific terms, general terms and negative specific terms based on their appearances in a training set.

In RFD model we introduce the term weighting method. Term coverage$^+$ is the set of relevant documents that contain t, and its coverage$^-$ is the set of irrelevant documents that contain t. We assume that the terms frequently used in both relevant documents and irrelevant documents are general terms. Therefore, we want to classify the terms that are more frequently used in the relevant documents into the positive specific category; the terms that are more frequently used in the irrelevant documents are classified into the negative specific category.

RFD model included three components. The first component analyzed the semantic structure of sentences; the second component constructed a conceptual ontological graph (COG) to describe the sematic structures; and the last one is find top concepts according to the first two components to generate feature vectors by using the standard vector space model.

This method is used for finding relevance features into three approaches. The first approach tries to diminish weights of terms that appear in both relevant documents and irrelevant documents. The second one is based on how often features appear or do not appear in relevant and irrelevant documents. The third one is based on finding features through positive patterns. RFD model used irrelevantdocuments in the training set in order to remove the noises.

IV. CONCLUSION

In this paper RFD model uses a feature clustering technique to automatically group terms into the three categories: positive specific features, general features, and negative specific features. The first issue in using irrelevant documents is how to select a suitable set of irrelevant documents since a very large set of negative samples is typically obtained. For example, a GoogleSearch can return millions of documents; however, only a few of those documents may be of interest to a Web user. Obviously, it is not efficient to use all of the irrelevant documents.This model is a supervised approach that needs a training set including both relevant documents and irrelevant documents.It also provides recommendations for offender (irrelevant) selection and the use of specific terms and general terms for describing user information needs. This model finds both positive and negative feedback and the RFD used irrelevant documents in the training set in order to remove the noises and also it can achieve the satisfactory performance.

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.

DOI 10.17148/IJARCCE.2015.4993

[4] N. Azam and J. Yao, "Comparison of term frequency and document frequency based feature selection metrics in text categorization," Expert Syst. Appl., vol.3 9, no. 5, pp. 4760–4 768, 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] N.Zhong, Y.Li, and S.-T. Wu, "Effective pattern discovery for textmining," in IEEE Trans. Knowl. Data Eng., vol. 24, no. 1, pp. 30–44, Jan. 2012.

[9] 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.

[10] 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.

[11] 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.

DOI 10.17148/IJARCCE.2015.4993