An Improved Model of Relevance Feature Discovery for Text Classification

M.H.Abuthakeer
Assistant Professor(Sl.Gr)/IT, Velalar College of Engineering and Technology,Thindal

E.Sowmiya,S.Padmavathi
IV-B.Tech IT, Velalar College of Engineering and Technology, Thindal

Abstract: The quality of discovered relevance features in text documents for describing user preferences cannot be guaranteed easily.The existing systems used the pattern and term based approach with different models such as Pattern Taxonomy Mining(PTM),Concept Based Model(CBM) etc.,The main challenge in the existing systems is integration of both terms and pattern features together and also it suffered from polysemy and synonymy.The Relevance Feature Discovery(RFD) comes as a breakthrough to the above disadvantages.The RFD model discovers both positive and negative terms from text documents and classifies them into categories and updates term weights.The Relevance Feature Discovery (RFD) is to find the useful features available in the text documents including both the relevant and irrelevant ones for describing the text mining results.

I. INTRODUCTION

The search engines retrieve a large amount of data according to the user preferences.It may contain both the relevant and irrelevant documents.The objective of Relevance Feature Discovery (RFD) is to find the useful features available in the text documents including both the relevant and irrelevant ones for describing the text mining results.The user submits a query and the search engines retrieve many documents according to the query submitted.The user analyses the documents and provides the feedback such as D+ for relevance and D- for irrelevance.This is known as the Relevance Feedback.The idea of Relevance Feedback(RF) is to involve the user in the retrieval process.

II. LITERATURE SURVEY

Relevance feature discovery for text analysis

The quality of discovered relevant features in text documents according to the user preferences is a big challenge to gurantee as there are so many terms,patterns and noise.The Relevance feature discovery solves this challenging issue by discovering both the positive and negative patterns in text documents as high level features in order to accurately weight low-level features based on their specificity and their distributions in the high-level features.

Effective pattern discovery for text mining:

The many data mining techniques have been proposed for mining useful patterns in text documents.The main issues is that how to effectively use and update discovered patterns in the domain of text mining.So an innovative and effective pattern discovery techniques which includes the processes of pattern deploying and pattern evolving to improve the effectiveness of using and updating discovered patterns for finding relevant and needed information.The operations involved are pattern mining,pattern evolving and information filtering.

Mining positive and negative patterns for relevance feature discovery:

It is a big challenge to clearly identify the boundary between positive and negative streams for information filtering systems. Several attempts have used negative feedback to solve this challenge; however, there are two issues for using negative relevance feedback to improve the effectiveness of information filtering. The first one is how to select constructive negative samples in order to reduce the space of negative documents. The second issue is how to decide noisy extracted features that should be updated based on the selected negative samples.

III. EXISTING SYSTEM

The relevance feature discovery is to find the useful features available in text documents including both relevant and irrelevant ones.There are two challenging issues in finding those patterns.They are the low-support problem and the misinterpretation problem.The former problem is that,long patterns are usually more specific

but they appear in the documents with low support or frequency. The latter comes with that,a highly frequent pattern may be frequently used in both relevant and irrelevant documents.The difficulty is how to use the discovered patterns to accurately weight useful features.The existing models such as the Pattern Taxonomy Mining(PTM) and Concept Based Model(CBM) solves the two challenging issues.The Pattern Taxonomy mining involves mining the closed sequential patterns in text paragraphs and deploying them over the term space.It splits all the text documents into paragraphs and it uses the frequent and closed patterns for pattern taxonomy mining.The concept based mining is used to discover the concepts by using the natural language processing.Feature Selection technique is also used for text classification and information filtering.The feature selection uses Bag-of-words technique.Many classifiers such as Naive Bayes,Rocchio,SVM have been developed but how to effectively integrate patterns in both relevant and irrelevant documents is still an open problem.

IV. Proposed Work

The proposed work involves an innovative technique for finding and classifying the low level terms based on their appearances in the high level features and their specificity in the training set.It also introduces a method to select the irrelevant documents that are closed to the extracted features in the relevant documents in order to effectively revise term weights.The proposed model has three major steps.They are feature discovery and deploying,term classification and term weighting.The RFD model describes the relevant features into three groups such as positive specific terms,general specific terms and negative specific terms.Here a term's specificity is defined according to its appearance in a given training set.The FClustering(Feature Clustering) categorizes the terms into positive terms(T+),general terms(G) and negative terms(T-) and groups them into clusters.The algorithm WFeature is applied to calculate term weights and then they are classified using FClustering algorithm.At last it chooses the first cluster as T+,second cluster as G and the last cluster as T-.
The contributions of the proposed model are,

1. It effectively uses the both relevant and irrelevant feedback to find useful features.
2. It integrates both term and pattern features together rather than using them in two separated stages.

V. 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 their appearances in the high-level patterns and the specificity. It also introduces a method to select irrelevant documents for weighting features.The RFD model also proves that the term classification can be done effectively by Feature Clustering method.The improved model automatically groups the terms into clusters.It provides a promising methodology for developing effective text mining models for relevance feature discovery.

REFERENCES

[1] Yuefeng Li,Abdulmohsen Algarni,Mubarak Albathan,Yan Shen and Moch Arif Bijaksana "Relevance Feature Discovery For Text mining",vol.6 ,June 2015.

[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] Y. Li, A. Algarni, and Y. Xu, "A pattern mining approach for information filtering systems," in Inf. Retrieval, vol. 14, pp. 237–256, 2011.

[6] 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.

[7] 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.

[8] S. Quiniou, P. Cellier, T. Charnois, and D. Legallois, "What about sequential data mining techniques to identify linguistic patterns for stylistics?" in Computational Linguistics and Intelligent Text Processing. New York, NY, USA: Springer, 2012, pp. 166–177.

[9] S.-T. Wu, Y. Li, and Y. Xu, "Deploying approaches for pattern refinement in text mining," in Proc. IEEE Conf. Data Mining, 2006, pp. 1157–1161.

[10] S.-T. Wu, Y. Li, Y. Xu, B. Pham, and P. Chen, "Automatic patterntaxonomy extraction for web mining," in Proc. Int. Conf. Web Intell., 2004, pp. 242–248.

[11] S. Shehata, F. Karray, and M. Kamel, "Enhancing text clustering using concept-based mining model," in Proc. 2nd IEEE Conf. Data Mining, 2006, pp. 1043–1048