Effective Pattern Discovery for Text Mining using Neural Network Approach

Harpreet Kaur[1], Rupinder Kaur[2]
[1, 2]*DIET, Kharar, Punjab, INDIA*

Abstract: It is observed that the text mining using the pattern discovery normally uses only the text material in normal fonts i.e. it does not consider the bold, underlined or italic or even the larger fonts as the key text pattern for text mining. This creates problem many a times when the key words are extracted from the document by the algorithm itself. In that case, important keywords are left from the main stream of text patterns.

In the proposed work, patterns are mined in both positive and negative feedback. It then automatically classifies the patterns into clusters to find relevant patterns as well as eliminate noisy patterns for a given topic. A novel pattern deploying strategy is proposed to extract high-quality features of text documents and use them for improving the retrieval performance. The proposed approach is evaluated by extracting features from RF to improve the performance of information filtering (IF).

Keywords: *Relevance Feedback, Text Mining*

I. INTRODUCTION

Many data mining techniques have been proposed for mining useful patterns in text documents. However, how to effectively use and update discovered patterns is still an open research issue, especially in the domain of text mining. Since most existing text mining methods adopted term-based approaches, they all suffer from the problems of polysemy and synonymy. Over the years, people have often held the hypothesis that pattern (or phrase)-based approaches should perform better than the term-based ones, but many experiments do not support this hypothesis.

It is difficult for users to precisely express their information needs for what they want. Relevance feedback (RF) is a promising approach that allows finding interesting or useful features from a training set to describe user information needs. Usually RF includes both positive (relevant) and negative (non-relevant) examples given by an individual user. The objective of Relevance Feature Discovery (RFD) is to find useful features available in the training set to describe user interests or preferences information. RFD is also of central interest in Information Retrieval (IR) and Data Mining (DM) communities. In IR, RFD is a basic step for building retrieval models, which find features from RF of text documents to improve retrieval performance. Most existing RF methods in IR are based on term-based approaches which use keywords or terms as features. The popular term-based methods include TF-IDF. The main advantages of the term-based algorithms include efficiently computational performance and mature theories for term weighting.

II. RELATED WORKS

In 2010, Taeho Jo published paper "NTC (Neural Text Categorizer): Neural Network for Text Categorization".This research proposes a new neural network for text categorization which uses alternative representations of documents to numerical vectors. Since the proposed neural network is intended originally only for text categorization, it is called NTC (Neural Text Categorizer) in this research.

In 2013 MICHAEL S. GASHLER, MICHAEL R. SMITH, RICHARD MORRIS, TONY MARTINEZ presented paper "Missing Value Imputation with Unsupervised Back propagation". They concluded that many data mining and data analysis techniques operate on dense matrices or complete tables of data. Real world data sets, however, often contain unknown values. Even many classification algorithms that are designed to operate with missing values still exhibit deteriorated accuracy.

In 2010, Taiwo Ayodele, Shikun Zhou, Rinat Khusainov presented paper "Email Classification Using Back Propagation Technique" This paper proposes a new email classification model using a teaching process of multi-layer neural network to implement back propagation technique.

In 2007, M. Govindarajan, and R. M. Chandrasekaran presented paper "Classifier Based Text Mining for Neural Network" in this paper they cocncluded Text Mining is around applying knowledge discovery techniques to unstructured text is termed knowledge discovery in text (KDT), or Text data mining or Text Mining. In Neural Network that address classification problems, training set, testing set, learning rate are considered as key tasks.

In 2012, Ning Zhong, Yuefeng Li, and Sheng-Tang Wu in their paper "Effective Pattern Discovery for Text Mining" they concluded that Many data mining techniques have been proposed for mining useful patterns in text documents. However, how to effectively use and update discovered patterns is still an open research issue, especially in the domain of text mining. Since most existing text mining methods adopted term-based approaches, they all suffer from the problems of polysemy and synonymy.

In 2012, Vandana Korde , C Namrata Mahender presented paper "TEXT CLASSIFICATION AND CLASSIFIERS: A SURVEY" .In this paper they are tried to give the introduction of text classification, process of text classification as well as the overview of the classifiers and tried to compare the some existing classifier on basis of few criteria like time complexity, principal and performance

In 2010 Aurangzeb Khan, Baharum Baharudin, Lam Hong Lee*, Khairullah khan published paper "A Review of Machine Learning Algorithms for Text-Documents Classification" The aim of this paper is to highlight the important techniques and methodologies that are employed in text documents classification, while at the same time

making awareness of some of the interesting challenges that remain to be solved, focused mainly on text representation and machine learning techniques. This paper provides a review of the theory and methods of document classification and text mining, focusing on the existing literature

III. ALGORITHM

Low-level features (i.e., terms) based on frequent patterns. However, this method originally evaluates the term weights of documents based on their distributions in all closed patterns, which include many general ones with high frequency. Such patterns often reduce the correctness of term weights. To eliminate the interference by the general patterns, weighting terms are performed by using all specific patterns (i.e., specific patterns and weak patterns).

Once relevant features are extracted, it is important to promote the weights to features which are specific for the positive documents to increase the discriminative power as well as decrease the weights of general features occurring in both the positive and negative documents.

To easily update the term weights, all the extracted features are portioned into either specific features ST or general features GT based on their appearances in specific patterns SP and weak patterns WP.

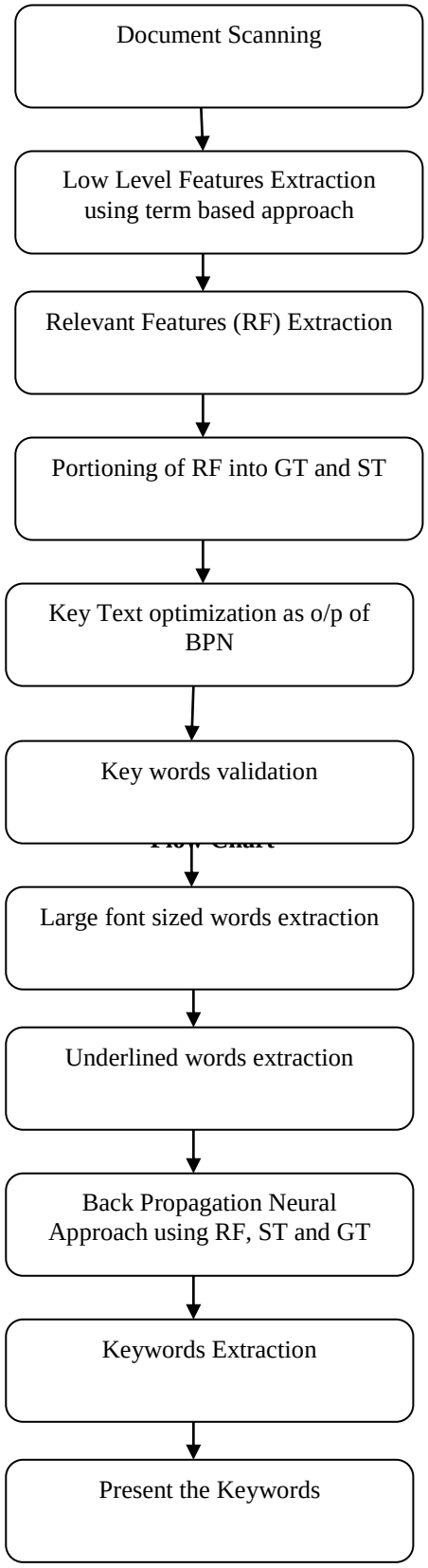
For each assessor topic, its data collection is split into two sets: a training set and a test set. All the meta-data information are removed and perform a common basic text processing for all documents, including stop-words removal according to a given stop-words list and stemming terms.

IV. BACK PROPAGATION NEURAL NETWORK

Back propagation is a learning technique that adjusts weights in the NN by propagating weight changes backward from the sink to the source nodes. Back propagation is the most well know form of learning because it is easy to understand and generally applicable. Back propagation can be thought of as a generalized delta rule approach. During propagation, data values input at the input layer flow through the network, with final values coming out of the network at the output layer. The propagation occurs by applying the activation function at each node, which then places the output value on the arc to be sent as input to the next nodes. In moat cases, activation function produces only one output value that is propagated to the set of connected nodes. The NN can be used for classification and/or learning. During the classification process, only propagation occurs. However, when learning is used after the output of the classification occurs, a comparison to the known classification is used to determine how to change the weights in the graph. In the simplest types of learning, learning progresses from the output layer backward to the input layer. Weights are changed based on the changes that were made in weights in subsequent arcs. The backward learning process is called back propagation.

V. RESULTS AND CONCLUSION

The presented technique is based on back propagation neural network and relevance feedback. It automatically classifies the patterns into clusters to find relevant patterns as well as eliminate noisy patterns for a given topic. A novel pattern deploying strategy is proposed to extract high-quality features of text documents and use them for improving the retrieval performance. Once relevant features are extracted, it is important to promote the weights to features which are specific for the positive documents to increase the discriminative power as well as decrease the weights of general features occurring in both the positive and negative documents.



VI. REFERENCES

[1] Raymond Chan , Qiang Yang , Yi Dong Shen "Mining High Utility Itemsets" Proceedings of the Third IEEE International Conference on Data Mining (ICDM'03) 0-7695-1978-4/03 $ 17.00 © 2003 IEEE.

[2] M. Govindarajan, and R. M. Chandrasekaran "Classifier Based Text Mining for Neural Network" World Academy of Science, Engineering and Technology International Journal of Computer, Information Science and Engineering Vol:1 No:3, 2007

[3] Aurangzeb Khan, Baharum Baharudin, Lam Hong Lee*, Khairullah khan "A Review of Machine Learning Algorithms for Text-Documents Classification" JOURNAL OF ADVANCES IN INFORMATION TECHNOLOGY, VOL. 1, NO. 1, FEBRUARY 2010

[4] Taeho Jo School of Information Technology & Engineering Ottawa University Ontario, Canada "NTC (Neural Text Categorizer): Neural Network for Text Categorization" International Journal of Information Studies Volume 2 Issue 2 April 2010.

[5] Taiwo Ayodele, Shikun Zhou, Rinat Khusainov "Email Classification Using Back Propagation Technique" International Journal of Intelligent Computing Research (IJICR), Volume 1, Issue 1/2, March/June 2010

[6] Ning Zhong, Yuefeng Li, and Sheng-Tang Wu "Effective Pattern Discovery for Text Mining" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 1, JANUARY 2012

[7] Vandana Korde , C Namrata Mahender "TEXT CLASSIFICATION AND CLASSIFIERS: A SURVEY International Journal of Artificial Intelligence & Applications (IJAIA), Vol.3, No.2, March 2012

[8] MICHAEL S. GASHLER MICHAEL R. SMITH, RICHARD MORRIS, TONY MARTINEZ "Missing Value Imputation With Unsupervised Backpropagation" arXiv:1312.5394v1 [cs.NE] 19 Dec 2013

[9] Luepol Pipanmaekaporn "Feature Discovery in Relevance Feedback Using Pattern Mining" 978-1-4799-0174-6/13/$31.00 ©2013 IEEE

[10] Nitu Mathuriya, Dr. Ashish Bansal " Comparison of K-means and Backpropagation

[11] Data Mining Algorithms" International Journal of Computer Technology and Electronics Engineering (IJCTEE) Volume 2, Issue 2

She is pursuing her M.Tech. in CSE from DIET, Kharar, Punjab India. Her field of interest is in software engineering and efforts estimation.