International Journal On Engineering Technology and Sciences – IJETS™
ISSN(P): 2349-3968, ISSN (O): 2349-3976
Volume III, Issue V, May- 2016

AN EFFICIENT FEATURE DISCOVERY AND TERM CLASSIFICATION USING SVM ALGORITHM

Nisha Ranjani.S[1] Karthikeyan.K[2]
PG Scholar[1], Assistant Professor[2]
Department of Computer Science & Engineering,
SNS College of Engineering,
Coimbatore.

ABSTRACT

Text mining, also known as text data mining, it refers to the process of deriving the quality information from the text. To guarantee relevance feature quality in text document is a big challenge because the text document having data patterns large scale terms. The existing text mining approaches are suffered from two major problems termed as polysemy and synonymy. In the proposed method some text preprocessing methods are applied in different dataset. It also reduces the features and updates the weight for each term to enhancing the text classification accuracy. Further training the classifier by Naive Bayesian (NB) and Support vector machine (SVM) algorithms to find the efficiency of the classification.

Key words: Feature selection; Text classification; Relevance feature.

I INTRODUCTION

Text mining has a large range of applications such as text summarization, categorization, entity and sentimental analysis. Text mining requires pre-processing which the text must be decomposed into smaller units such as terms and phrases. For example, in some text mining applications, terms extracted from the documents and treated as features. Text clustering is also termed as document clustering. Relevance feature discovery (RFD) is a classical, but challenging task in IR and text mining. The objective of RFD is to find useful features in user relevance feedback (typically text documents) to fulfill user information needs. Traditionally, relevance feedback has been used widely in the area of IR to improve search quality corresponding a given query. It has been currently used in text mining systems. For example, information filtering (IF) and text classification. Relevance feedback is a subset of retrieved documents that have been judged by the users. The judgment can be 1 which means it is related to the user's topic of interest or 0 which means it is not related to what users want.

II EXISTING METHODS

1. COMPARISON OF TERM FREQUENCY AND DOCUMENT FREQUENCY BASED FEATURE SELECTION METRICS IN TEXT CATEGORIZATION

Text categorization plays an important role in applications where information is filtered, monitored, personalized, categorized, organized or searched. Feature selection remains as an effective and efficient technique in text categorization. Feature selection metrics are commonly based on term frequency or document frequency of a word. Experimental results revealed that the term frequency based metrics may be useful especially for smaller feature sets.

2. TOPICAL PATTERN BASED DOCUMENT MODELLING AND RELEVANCE RANKING

Topic modelling was proposed to generate statistical models to represent multiple topics in a collection of documents, but in a topic model, topics are represented by distributions over words which are limited to distinctively represent the semantics of topics. This paper proposes a novel information filtering model, Significant matched Pattern-based Topic Model (SPBTM).

3. MINING POSITIVE AND NEGATIVE PATTERNS FOR RELEVANCE FEATURE DISCOVERY

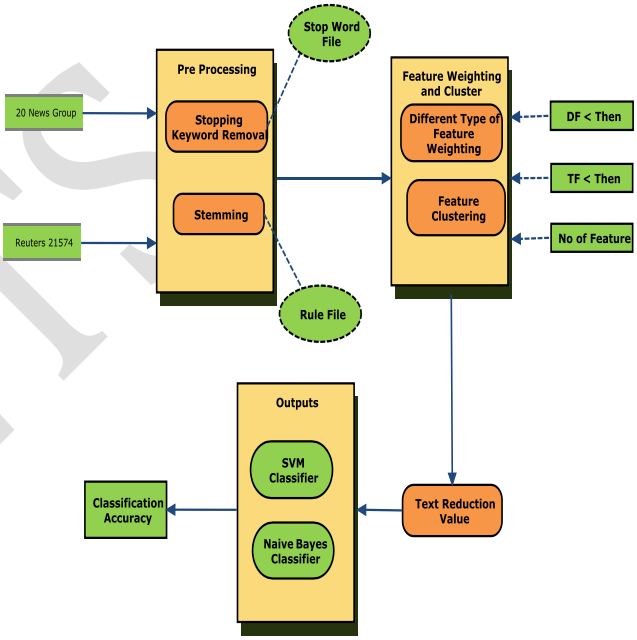
Most existing popular text mining and classification methods have adopted term-based approaches. However, they have all suffered from the problems of polysemy and synonymy. The innovative technique presented in paper makes a breakthrough for this difficulty. This technique discovers both positive and negative patterns in text documents as higher level features in order to accurately weight low-level features (terms) based on their specificity and their distributions in the higher level features.

4. A FAST CLUSTERING-BASED FEATURE SUBSET SELECTION ALGORITHM FOR HIGH-DIMENSIONAL DATA

A feature selection algorithm may be evaluated from both the efficiency and effectiveness points of view. 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 a subset of features.

III PROPOSED WORK

In the proposed method, machine learning methods for text classification is used to apply some text preprocessing methods in different dataset, and then to extract a feature vector construction for each new document by using feature weighting and feature selection algorithms for enhancing the text classification accuracy. Further training the classifier by Naive Bayesian (NB) and Support vector machine (SVM) algorithms so that,the efficiency of the classification method can be findout.



IV MODULES

1. Data Preprocessing
2. Feature weighting and reduction
3. Text classification

Data preprocessing:

It is a data mining technique that involves transforming of raw data in to understandable formats. The dataset are taken and they are preprocessed by using the stemming algorithm. Stemming or lemmatization is a technique for the reduction of words into their root.there are two

algorithm used for preprocessing the documents. It includes,

- Porter stemming
- Lancaster

Feature weighting and reduction:

In feature selection process the weighting function is selected and with the method of summation the weighting function includes **odds ratio, information gain, document frequency , term frequency, mutual information, chi square.** The number of features are reduced according the user and the dimensionality are reduced . Now the reduced features are weighted for each term with the term frequency and document frequency with their inverse frequency.

Term classification:

To know the efficiency the classification SVM and naïve Bayesian algorithms are used. **"Support Vector Machine"** (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate.

V.CONCLUSION

In this paper, each feature (term or single word) is assigned with a score according to a score-computing function. Then those with higher scores are selected.. Classification includes different parts such as text processing, feature extraction, feature vector construction and final classification. Here first try to apply some text preprocess in different dataset, and then extract a feature vector for each new document by using feature weighting and feature selection algorithms for enhancing the text classification accuracy. After that train our classifier by Naïve Bayesian (NB) and support vector machine (KNN) algorithms. In Experiments, although both algorithms show acceptable results for text classification.

VI.REFERENCES

[1] Azam.N and Yao.J, "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.

[2] Chandrashekar .G and Sahin.F, "Asurvey on feature selection methods," in Comput. Electr. Eng., vol. 40, pp. 16–28, 2014.

[3] Daniel Engel, Lars Hüttenberger, Bernd Hamann, A Survey of Dimension Reduction Methods for High-dimensional Data Analysis and Visualization, LNCS Springer, 2014, pp. 1-16.

[4] Khalid, Samina, Khalil Tehmina, Nasreen Shamila, A Survey of Feature Selection and Feature Extraction Techniques in Machine Learning, IEEE Science and Information Conference, 2014, pp. 372-378.

[5] Li.Y, Hus.D.F, and ChungS.M, "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.

[6] Ranshul Chaudhary, Prabhdeep Singh, Rajiv Mahajan, *A Survey on Data Mining Techniques*, International Journal of Advanced Research in Computer and Communication Engineering, Volume 3, Issue 1, 2014, pp. 5002-5003.

[7] Salton.G and Buckley.C, "Term-weighting approaches in automatic text retrieval," in Inf. Process. Manage., vol. 24, no. 5, pp. 513–523, Aug. 1988.

[8] Song.Q, Ni.J, and Wang.G, "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.

[9] Yang.Y and , Pedersen.J.O, "A comparative study on feature selection in text categorization," in Proc. Annu. Int. Conf. Mach.Learn., 1997, pp. 412–420.