PATTERN BASED CLASSIFICATION FOR TEXT MINING USING FUZZY SIMILARITY ALGORITHM

M.Janani, N.Vennila, R.Swathi

Abstract— Text mining is the process of extracting of useful information from structured and unstructured documents. There are many methods which have been proposed for text mining used in pattern in text documents. Text mining mainly concentrated to detecting the different entities such as word, phrases, term, pattern, concept, paragraph, sentence, & documents. The system assigns the frequency to each and every word, the weight of all the document is utilized for pattern clustering. Pattern clustering is one of the most used methods for feature extraction in text classification. In this paper propose a pattern based classification for text mining using fuzzy similarity algorithm. It overcomes the low frequency problem & misinterpretation, and also estimates the similarity between the different patterns and word in effective manner.

I. INTRODUCTION

A novelty and effective pattern discovery technique which includes the processes of pattern deploying and pattern evolving, to improve the effectiveness of using and updating discovered patterns for finding relevant and interesting information [1]. The aim of pattern clustering is grouping the entire original feature into clusters. This propose a fuzzy estimated and similarity - based self generating algorithm for text classification. It allows the precise text representation, and opposes the incomplete data, distinct due to different text categorization. It is used to reduce the data sets, redundant data & not clearly understood data. These approaches are precisely classifying text document up-to 80%.Anovel coclustering algorithmnamed Locally Discrimin ative Coclustering (LDCC) to explore the relationship between samples and features as well as the inter sample and inter feature relationships [2]. New discretization technique EDISC which utilizes the entropy-based principle but takes a class-tailored approach to discretization [3]. An analysis the problem degrades the quality of the clustering result, a new link-based approach, which improves the conventional matrix by discovering unknown entries through similarity between clusters in an ensemble [4]. The proposed ensemble classifier is based on original concepts such as learning of cluster boundaries by the base classifiers and mapping of cluster confidences to class decision using a fusion classifier [5].

Manuscript received Mar 02, 2015

M.Janani, Department of Information Technology, Jeppiaar Engineering College, Chennai, India

N.Vennila, Department of Information Technology, Jeppiaar Engineering College, Chennai, India

R.Swathi, Department of Information Technology, Jeppiaar Engineering College, Chennai, India

A two-stagemulti-feature word-level Chinese paraphrase extracting method. In stage one, using data mining in stage two, stratified probability statistical model is established [6]. The proposed two pattern refinement methosto deploy the discovered patterns into a feature space which is used to represent the concept of documents. Our methods adopt the mining sequential pattern technique to find semantic patterns from text documents and then deploy these patterns using proposed deploying algorithms [7]. Most of the text is represented by following format such as word, phrases, phrases, term, pattern, concept, paragraph, sentence, & documents. Main issues occur in text mining such as low frequency problem, misinterpretation problem. These methods include pattern taxonomy, concept based model, association mining, relevance feature discovery, and iterative learning algorithm was proposed. These approaches have shown some improvements in text mining. This paper illustrates a new model which integrates topic filtering and pattern taxonomy mining together to alleviate information overload and mismatch problems. The proposed method has been evaluated using the standard TREC routing framework [8].This paper presents an innovative approach for relevance feature discovery. It introduces a method to select negative documents (or called offenders) that close to the extracted features in the positive documents. It also proposes an approach to revise low level features (terms) based on both their appearances in the higher level features (patterns) and their categories (the positive specific category, general category and negative specific category [9].

II. PROPOSED SYSTEM

Most of the text is represented by following format such as word, phrases, phrases, term, pattern, concept, paragraph, sentence, & documents. Main issues occur in text mining such as low frequency problem, misinterpretation problem. These methods include pattern taxonomy, concept based model, association mining, relevance feature discovery, iterative learning algorithm were proposed. These approaches have shown some improvements in text mining. we propose a pattern based classification in text mining using a fuzzy similarity algorithm. It overcomes the low frequency problem & misinterpretation, and also estimates the similarity between the different patterns and word in an effective manner. Most of the text is represented by following format such as word, phrases, phrases, term, pattern, concept, paragraph, sentence, & documents. Main issues occur in text mining such as low frequency problem, misinterpretation problem. These methods include pattern taxonomy, concept based model, association mining, relevance feature discovery, iterative learning algorithm were proposed. These approaches have shown some improvements in text mining.

III. SYSTEM ARCHITECTURE

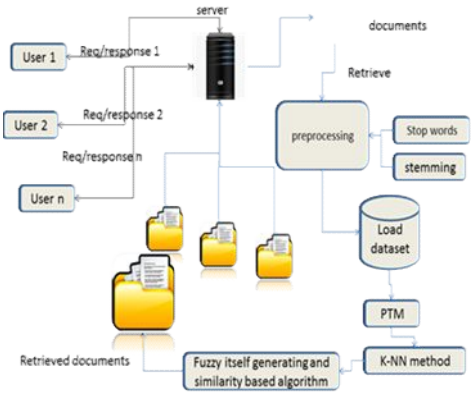

Figure (a)

This a client and server environment where the client gives a request and server process the request and sent a response .we can have n number of clients for one server so, in the above figure as we assumed N user client with one server . After the query is processed from the server to the textual documents, it starts to preprocess. Preprocessing means reduce the training time, storage requirements and mainly used to remove the unwanted words from the textual documents. In this preprocessing consists of two operations such as stemming, stop list. Stemming means queue words having same meaning. And stop words means irrelevant word or root word. After the process is completed in preprocessing and then load the dataset. After that the next step is pattern taxonomy. So the extracted sets of text documents go to the Pattern taxonomy, which is used to improve the performance of closed pattern. It can extract the word from each and every document. In this PTM , discovering the pattern for calculating the specificities of the particular pattern in textual documents, and minimize the problem such as misinterpretation, low frequency problem which one reduce the noisy patterns. So it can improve the accuracy by two processes such as pattern deploying (PD) method & inner pattern evolving (IPE). Pattern deploying method does discover the pattern in efficient manner and it's easily deletes overlapping among the pattern. And then inner pattern deploying does reduce the side effects of noisy patterns, low frequency problem. Next step is K-nearest neighbor method and then fuzzy similarity it is used to extract the index word or term from particular textual documents. Finally the retrieved documents are going to the client through the server.

IV. INFORMATION RETRIVEL

Stemmers are common in query analysis systems such as Web search engines. The effectiveness of word stemming for English query systems was soon found to be limited, however, and this has led early information retrieval researchers to deem stemming irrelevant in general.

Domain Analysis

Stemming is used to determine domain vocabularies in domain analysis.

Use in commercial products

Many commercial companies have been using stemming since at least the 1980s and have produced algorithmic and lexical stemmers in many languages. The Snowball stemmers have been compared with commercial lexical stemmers with varying results. Google search adopted word stemming. Before adoption of word stemming a search for "fish" would not have returned "fishing". Other software search algorithms vary in their use of word stemming. Programs is simply to search for a substrings obviously will find "fish" in "fishing" but when searching for "fishes" will not find occurrences of the word "fish". where the output of PTM is given to K-NN method Where K-NN stands for k-Nearest Neighbors algorithm To fuzzy itself generating and similarity based algorithm and final the document is retrieved by the server after all the above process is done. KNN stands for K-Nearest Neighbors it is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure it's also known as K-Nearest Neighbors, Memory-Based Reasoning, Example-Based Reasoning, Instance-Based Learning, Case-Based Reasoning, Lazy Learning *it's applicable in various fields like* Classification and Interpretation legal, medical, news, banking Problem-solving planning, pronunciation Function learning dynamic control Teaching and aiding help desk, user training KNN is conceptually simple, yet able to solve complex problems Can work with relatively little information Learning is simple (no learning at all!) Memory and CPU cost Feature selection problem Sensitive to representation

V. PROPOSED METHOD

The proposed model is an extension of the work. The proposed pattern mining model consists of pattern deploying, inner pattern evolution, and fuzzy estimated and similarity based self generating algorithm for text classification. It is one of the extended fuzzy feature clustering algorithms in text Classification. A textual document is input to the proposed system. It is collecting from different sectors such as news papers, articles, journals, magazine, university details and IT industry. Here we can extract the word from text document. Same words are classified into one group. Different words are categorized into another one group. After classification we calculate similarity between different words. Then each word has been weights in a desired number of categorizations are formed automatically. Four ways of weighting, paragraph, document, concept and sentence are discussed. By fuzzy algorithm can properly deals with the weighting scheme between different features. It is very easy to avoiding the low frequency and misinterpretation problem in pattern mining. We conduct some real world experiments on data sets RCV1 and TREC. It become show excellent improvements and also run faster, reduces storage requirements. Pattern reduction. There are two ways of doing pattern reduction, pattern selection, pattern extraction. By pattern selection approaches, a new pattern set p'= {p'1, p'2… p''k} is carried. Which is a sub set of the initial pattern set p, where is represented by following format p= {p1, p2… p k}. Information gain due to a particular splits of pattern p into pi, i= 1, 2. . . k is then information gain(p,{ p1, p2… pk}) = purity (p) - purity (p)1, p2,……, pk) , it measures uncertainty of pattern in effective manner, then which can assign the weight to each and every pattern in a given set of text groups. Pattern clustering pattern or word clustering is an extended approach for feature reduction. Which can used to categorizes the all words into some

groups , here we consider as two kinds Cluster group such as inter cluster, intra cluster. Inter cluster means all words present in within Boundary. Intra cluster means all words or features present in outside of the boundary.

CONCLUSION

This paper proposed data mining techniques and also different text classification approach. These methods incorporate pattern taxonomy model, K- nearest neighbor method, concept based model in the last decade. Pattern clustering is one of the great techniques for feature extraction in text classification. The principle point of pattern clustering is to gathering the all unique feature into clusters. In this paper we propose an approach. These methodologies are precisely grouping content archive, attain to 80% of precision. Yet we are expecting our proposed strategy is arriving at the focus over 90%. The concept based model (CBM) has likewise enhancing the better execution. Fluffy evaluated and closeness based self producing calculation for content arrangement. It is beat the low recurrence issue, additionally compute the comparability between the diverse content records in compelling, exact way. Words are sorted into two types of group gatherings. One is cluster group (comparable), another is intra cluster group (unique). Exploratory on RCV1 data collection and TREC points execute that the proposed result attains to better execution. This study is as yet having changes on it and particularly on the looking into results. In future, can execute the fluffy self producing and similitude based calculation utilizing java technology. As of now, the project as performed well up to the feasibility study of the proposed system. Proposed system will likewise incorporate assessing fuzzy estimated and comparability based self generating algorithm with different datasets, so that the guarantee of efficiency of the proposed system. There is additionally yet more extension for future research in the field of idea concept-based text classification and an alternate direction is to apply the same methodology to web document classification.

REFERENCE

[1] [Title Effect ve Pattern Discovery for the text mining Author: Ning Zhong, Yuefeng Li, Sheng-Tang Wu Year of publish: 2012
[2] Title Locally Discriminative Coclustering Author: Lijun Zhang Year of publish: 2012
[3] Title EDISC: A class-Tailored Discretization Technique for Rule-Based Classification Author: Khurram Shehzad Year of publish: 2012
[4] Title EDISCA link-Based Cluster Ensemble Approach for Categorical Data Clustering Author: Natthakan Iam-0n Year of publish: 2012
[5] Title Cluster-Oriented Ensemble Classifier: Imapct of Multicluster Characterization on Ensemble Classifier learning Author: Brijesh Verma Year of publish: 2012
[6] Title ClusterA research on Multi-feature word-Level Parapharase Extracting System Based on Context Author: Y.Li, X.Zhou, P.Bruza, Y.Xu and R.Y.Lau Year of publish: 2008
[7] Title Deploying Approaches for Pattern Refinement in Text Mining AuthorSheng-Tang Wu Yuefeng Li Yue Xu Year of publish: 2006
[8] Title A Two-stage Information Filtering Based on Rough Decision Rule and Pattern Mining Author: Xujuan Zhou*, Yuefeng Li*, Peter Bruza*, Yue Xu* and Raymond Lau Year of publish: 2010
[9] Title Mining Positive and Negative Patterns for RelevanceFeature Discovery Author: Yuefeng Li Abdulmohsen Algarni Ning Zhong Year of publish: 2010
[10] Title Mining Multi-Faceted Overviews of Arbitrary Topics in a Text Collection Author: Xu Ling, Qiaozhu Mei, ChengXiang Zhai, Bruce Schatz Year of publish: 2008.