An Efficient Pattern Discovery Over Long Text Patterns

T.Ravi Kiran[1], Ch.Sai Priyanka Rani[2], K.Bhoomika[3], K.Madhuri[4], L.Naresh[5]
Assistant Professor[1],B.Tech Scholar[2,3,4,5]
Dept of CSE, VITS College of Engineering, Sontyam, Visakhapatnam, Andhra Pradesh

Abstract: **There are several techniques are implemented for mining documents. In this text mining, still so many problems getting exact patterns in text mining. In this some of the techniques are adapted in text mining. In proposed system the temporal text mining approach is introduced. The system terms of its ability is evaluated to predict forthcoming events in the document. In this we present optimal decomposition of the time period associated with the given document set is discovered where each subinterval consists of consecutive time points having identical information content. Extraction of sequences of events from new and other documents based on the publication times of these documents has been shown to be extremely effective in tracking past events.**

I. INTRODUCTION

Text mining is also called *text data mining* which is roughly equivalent to text analytics and refers to the process of deriving highquality data from text. Highquality data is typically derived through the devising of patterns and trends through means such as statistical pattern learning. Text mining includes the process of structuring the input text usually parsing along with the addition of some derived linguistic features and the removal of others and subsequent insertion into a database deriving patterns within the structured dataand finally evaluation and interpretation of the output.

Another common application for text mining is to aid in the automatic classification of texts. For example, it is possible to "filter" out automatically most undesirable "junk email" based on certain terms or words that are not likely to appear in legitimate messages however instead identify undesirable electronic mail and such messages can automatically be ignored. This automatic systems for classifying electronic messages can also be useful in applications where messages need to be routed to the most appropriate department or agency for example email messages with complaints or petitions to a municipal authority are automatically routed to the appropriate departments at the same time and the emails are screened for inappropriate or obscene messages and which are automatically returned to the sender with a request to remove the offending words or content.

In some business domains, the majority of information is collected in open-ended in textual form. Consider example warranty claims or initial medical interviews can be summarized in brief narratives or when you take your automobile to a service station for repairs the attendant will write some notes about the problems that you report and what you believe needs to be fixed. The notes are collected electronically and so those types of narratives are readily available for input into text mining algorithms. This information can then usefully exploited and for example identify common clusters of problems and complaints on certain automobiles. Open-ended descriptions by patients of their own symptoms might yield useful clues for the actual medical diagnosis.

As the volume of electronic information increases and there isgrowing interest in developing tools to help people better find the filter these resources. Text categorization isthe assignment of natural language texts to one or morepredefined categories based on their content which is animportant component in many information organization andmanagement tasks. Machine learning methodsincludes Support Vector Machines have tremendous potentialfor helping people to effectively organize the electronicresources. Text mining often involves the extraction ofkeywords with respect to some measure of importance.

Weblog data is textual content with a clear and significanttemporal aspect. Text categorization is the task of automaticallysorting a set of documents into categories from a predefinedset and this task has several applications including automatedindexing of scientific articles according to predefined thesauriof technical terms.Filing patents into patent directories and selective dissemination of information to informationconsumers.The population of hierarchical catalogues ofWeb resources and spam filtering identification of documentgenre and authorship attribution evenautomated essay grading. The text classification ismore effective because it organizations from the need ofmanually organizing document bases and which can be tooexpensive or simply not feasible given the time constraints ofthe application or the number of documents involved.

Theaccurate results in text classification systems rivals that oftrained human professionals and thanks to a combination ofinformation retrieval technology and machine learningtechnology. The outline the fundamental traits ofthe technologies involved that can feasiblybe tackled through text classification and of the tools andresources thatis available to the researcher and developerwishing to take up these technologies for deploying real-worldapplications. A web technology extracts the statisticalinformation and discovers interesting user patterns and cluster theuser into groups according to their navigational behavior thendiscover potential correlations between web pages and usergroups of identification of potential customers for E-commerce and enhance the quality and delivery of Internet informationservices to the end user to improve web server systemperformance and site design and facilitate personalization.

II.RELATED WORK

Before any classification task, one of the most fundamental tasks thataccomplished is that of document representation and feature selection. Until the feature selection is also desirable in other classification tasks and it is especially important in text classification due to the high dimensionality of text features and the existence of irrelevant features. The text can be represented in two separate solutions. The first is as a group of words in which a document is represented as a set of words and together with their associated frequency in the document. The representation is independent of the sequence of words in the collection. Next method is to represent text directly as strings which each document is a sequence of words. Many classification methods use the groupwords representation because of its simplicity for classification purposes. We will discuss some of the methods which are used for feature selection in text classification.

The most common feature selection which is used in both supervised and unsupervised applications is that of stop-word removal and stemming. We determine the common words in the documents which are not specific or discriminatory to the different classes. Different forms of the same word are consolidated into asingle word. Singular or plural and different tenses are consolidatedinto a single word. That these methods are not specificto the case of the classification problem and are often used in a varietyof unsupervised applications such as clustering and indexing. In thecase of the classification problem and it makes sense to supervise the featureselection process with the use of the class labels. This type of selectionprocess ensures that those features which are highly skewed towards the presence of a particular class label are picked for the learning process.

Information Gain

Another related measure which is commonly used for text feature selection is that of information gain or entropy. Let P_i be global probability of class I and $p_i(w)$ be the probability of class I that is given that the document contains the word w. Let F(w) be the fraction of documents containing the word w. The information gain measure w is defined as follows:

$$I(w)=-\sum_{i=1}^{k} Pi.\log(p_i)+F(w).\sum_{i=1}^{k} Pi(w).\log(P_i(w))+$$

$$+(1-F(w)).\sum_{i=1}^{k} Pi(w)).\log(1-p_i(w))$$

The greater the value of the information gain I(w), the greater the discriminatory power of the word w. For a document corpus containing n documents and d words, the complexity of the information gain computation is O(n · d · k).

χ2-Statistic

The χ2 statistic is a different way to compute the lack of independence between the word w and a particular class i. Consider n be the total number of documents in the collection and $p_i(w)$ be the conditional probability of class i for documents which contain w, P_i be the global fraction of documents contains the class i and F(w) be the global fraction of documents which includes the word w. The χ2- of the word between word w and class i is defined as follows:

$$X^2_i(w)=\frac{n.F(w)\text{^}2.(pi(w)-pi)\text{^}2}{F(w).(1-F(w)).pi.(1-pi))}$$

In this case we can compute a global χ2 statistic from the class specific values. And we use either the average of maximum values in order to create the composite value:

$$X^2_{avg}(w)=\sum_{i=1}^{k} pi.X_i^2(w)$$

$$X^2_{max}(w)=\max_i X_i^2(w)$$

χ2-statistic and mutual information are different ways of measuring the correlation between terms and categories. The main advantage of the χ2-statistic over the mutual information measure and results normalized valueand those values are more comparable across terms in the same category.

III.PROPOSED WORK

Pattern Methodology Model

A)Features Selection Method

Initially the documents are considered as an input and the features for the set of documents are collected. Features are selected based on the TFIDF method. Information retrieval

was developed based on many mature techniques which demonstrate the terms which are important features in the text documents. Many terms with larger weights consider the term frequency and inverse document frequency tf*idfweighting scheme are general terms because they can be frequently used in both relevant and irrelevant information.The selection features approach is used to improve the accuracy of evaluating term weights because the discovered patterns are more specific than whole documents. Many approaches have been conducted by the use of feature selection techniques.

Finding Frequent and Closed Sequential Pattern

When feature selection process is completed then frequent and closed patterns are discovered based on the documents and termset 'X' in document'd', ⌜X⌝ is used to denote the covering set of X for d and which includes all paragraph dp€ PS(d) such that
$X \subseteq d_p$

$X=\{d_p | d_p \in PS(d), X \subseteq d_p\}$

Its absolute support is the number of occurrences of X in PS(d) $SUP_a(X)=|X|$. Its relative support is the fraction of the paragraphs that contain the pattern$SUP_r(X)=|X|/|PS(d)|$ Patterns can be structured into methodology by using the subset relation. Small patterns in the methodology are usually general because they could be used frequently in both positive and negative documents and larger patterns are usually more specific since they may be used only in positive documents.

The semantic information will be used in the pattern process to improve the performance of using closed patterns in text mining. A sequential pattern X is called frequent pattern if its relative support is a minimum support. Some property of closed patterns can be used to define closed sequential patterns. The algorithm for finding the support count is given as,

Algorithm:

Input: Positive documents which contains more number of data, Negative documents which doesn't contains more number of data, minimum support.

Output: document patterns DP, and support of terms.

1. Document Patterns =ø
2. For d in Documents D do
    a. Consider PS(d) be the set of paragraphs in d,
    b. Sequential Patterns=SPMining(PS(d), min_sup);
    c. d'=ø
    d. for 'p' pattern € sequential pattern do

        $p=\{(t,1)|t€p_i$

        d'=d' XOR p

    e. Document Patterns=DP U {d'}

    end

3. T={t|(t,f) € p,p€DP}
4. For t € T
    a. Sup(t)=0

    end

5. For pattern in Document Patterns
    a. For t in Patterns

        Supp(t)=Supp(t)+w

    end

6. end

In PTM algorithm all the documents'd' are splitted into paragraphs P which yields PS (d). Patterns can be structured into methodology by using the relation (or subset). From the set of paragraphs in documents the frequent patterns and the covering sets are discovered for each. Smaller patterns in the methodology patterns are usually more general because they could be used frequently in both positive and negative documents. Larger patterns in the methodology are usually more specific since they may be used only in positive documents. The semantic information will be used in the pattern methodology to improve the performance of using closed patterns in text mining.

D-Pattern Discovery

D-pattern mining algorithm is used to discover the D-patterns from the set of documents. The efficiency of the pattern methodology mining is improved by proposing an SP mining algorithm to find all the closed sequential patterns and which is used as the well-known appropriate property in order to reduce the searching space. The algorithm describes the training process of finding the set of d-patterns. For every positive document and the SP Mining algorithm is first called giving rise to a set of closed sequential patterns. The main focus is the deploying process and it which consists of the d-pattern discovery and term support evaluation. All discovered patterns in a positive document are composed into a d-pattern giving rise to a set of d-patterns .The term supports are calculated based on the normal forms for all terms in d-patterns.

IV. CONCLUSION

Proposed model with new pattern discovery model for textmining mainly focuses on the implement of textpattern. We present adynamic programming algorithm for optimalinformation preserving decomposition and optimal decomposition is introduced. This is used for analyzingrelationship between the time periodassociated with the document set and the significantinformation computed for temporal analysis. It findspatterns quickly for various ranges of parameters. It focuses on usinginformation extraction to extract a structured database from acorpus of natural language text and then discover patterns and form of data-cleaning that identifiesequivalent but textually distinct items in the extracted dataprior to mining.

REFERENCES

1. M.F. Caropreso, S. Matwin, and F. Sebastiani.Statistical Phrases in Automated Text Categorization, Technical Report IEI-B4-07-2000, Instituto di Elaborazionedell'Informazione, 2000.
2. C. Cortes and V. Vapnik.Support-Vector Networks, Machine Learning, vol. 20, no. 3, pp. 273-297, 1995. 3. S.T. Dumais, Improving the Retrieval of Information from External Sources, Behavior Research Methods, Instruments, and Computers, Vol. 23, No. 2, pp. 229-236, 1991.
4. J. Han and K.C.-C.Chang.Data Mining for Web Intelligence, Computer, Vol. 35, No. 11, pp. 64-70, Nov. 2002.
5. J. Han, J. Pei, and Y. Yin.Mining Frequent Patterns without Candidate Generation, Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD '00), pp. 1-12, 2000.
6. Y. Huang and S. Lin. Mining Sequential Patterns Using Graph Search Techniques, Proc. 27th Ann. Int'l Computer Software and Applications Conf., pp. 4-9, 2003.7. N. Jindal and B. Liu.Identifying Comparative Sentences in Text Documents, Proc. 29th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '06), pp. 244-251, 2006.
8. T. Joachims. A Probabilistic Analysis of the Rocchio Algorithm with tfidf for Text Categorization, Proc. 14th Int'l Conf. Machine Learning (ICML '97), pp. 143-151, 1997.
9. T. Joachims. Text Categorization with Support Vector Machines: Learning with Many Relevant Features, Proc. European Conf. Machine Learning (ICML '98),, pp. 137-142, 1998.

BIOGRAPHIES

T.Ravi Kiran is an Assistant Professor in the Department of Computer Science & Engineering, VITS College of Engineering, Sontyam, Visakhapatnam, Andhra Pradesh. He has 5 years of experience in Teaching. His research interests include Cloud Computing, Web Technologies, Information Security, Data Mining, Search Engines, Information Retrieval, Network Security, Database Systems, Data Privacy, Image Processing, Computer Networks.

Ch.Sai Priyanka Rani is currently pursuing B.Tech. degree in Computer Science & Engineering, VITS College of Engineering, Sontyam, Visakhapatnam, Andhra Pradesh. Her research interests include Data Mining, Information Retrieval.

K.Bhoomika is currently pursuing B.Tech. degree in Computer Science & Engineering, VITS College of Engineering, Sontyam, Visakhapatnam, Andhra Pradesh. Her research interests include Data Mining, Information Retrieval.

K.Madhuri is currently pursuing B.Tech. degree in Computer Science & Engineering, VITS College of Engineering, Sontyam, Visakhapatnam, Andhra Pradesh. Her research interests include Data Mining, Information Retrieval.

L.Naresh is currently pursuing B.Tech. degree in Computer Science & Engineering, VITS College of Engineering, Sontyam, Visakhapatnam, Andhra Pradesh. His research interests include Data Mining, Information Retrieval.