International Journal of Trend in Scientific Research and Development (IJTSRD)
International Open Access Journal
ISSN No: 2456 - 6470 | www.ijtsrd.com | Volume - 2 | Issue – 3

An Efficient Pharse Based Pattern Taxonomy Deploying Method for Text Document Mining

[1] S. Brindha
Assistant Professor, Department of Computer Science Palanisamy College of Arts, Perundurai, Tamilnadu, India

[2] Dr. S. Sukumaran
Associate Professor, Department of Computer Science, Erode Arts and Science College, Erode, Tamilnadu, India

ABSTRACT

The extraction of multiple word which are related to expressions that has been increasingly a special topic in the last few years. Relevant expressions are applicable in diverse areas such as Information Retrieval, document clustering, or classification and indexing of documents. However, relevant single-words, which represent much of the knowledge in texts, have been a relatively dormant field. In this paper we present a statistical language-independent approach to extract concepts formed by relevant single and multi-word units. By achieving promising precision/recall values, it can be an alternative both to language dependent approaches and to extractors that deal exclusively with multi-words. In this paper proposed method pattern Taxonomy Deploying method to apply to find a new and efficient pattern method by which research related document, research related documents are patterned and classification of different field are done and more than 80% percent of the documents are successfully identified and categorized.

Keywords: *Pattern Taxonomy Deploying, Support Vector Machine, Pattern Taxonomy method*

I. INTRODUCTION

Text Mining (TM) field has gained a great deal of attention in recent years due the tremendous amount of text data, which are created in a variety of forms such as social networks, patient records, health care insurance data, research outlets, etc. The amount of text that is generated every day is increasing dramatically. This tremendous volume of mostly unstructured text cannot be simply processed and perceived by computers. Therefore, efficient and effective techniques and algorithms are required to discover useful patterns. Text mining is the task of extracting meaningful information from text, which has gained significant attentions in recent years. Text mining is the retrieving by computer machine of new, previously unknown information by automatically extracting information from different written text resources. Nowadays most of the text mining applications have established a grouping of research processing. A quantity of the applications is spam filtering, emails categorization, directory maintenance, ontology mapping, document retrieval, routing filtering etc. Text documents have become the most common container of information. Due to the increased popularity of the internet, emails, research group messages etc. The text is the dominant type of information to exchange. Many real times text mining applications have received a lot or research attention. Interacting with the web and with colleagues and friends to acquire information is a daily of many human beings. To acquire similar information on the web in order to gain specific knowledge in one domain. In a research lab, members are often focused on projects which require similar background knowledge. The classification problem assumes categorical values for the labels, though it is also possible to use continuous values as labels. This is referred to as the regression modeling problem. The problem of text classification is closely related to that of classification of records with set valued features. This model considered about the information about

the presence or absence of words is used in a document only. The problem of text mining and text classification finds application in a wide variety of domains in text mining. Some examples of domains in which text classification is commonly used. Mostly the research services are now a days are electronic in nature in which a large volume of Research articles are produced every solitary day by the organizations. In such cases, it is complicated to categorize the research articles manually. Therefore, automated related to methods can be very helpful for research categorization in a variety of web portals. This application is also referred to as text filtering.

Document organization and retrieval application is generally useful for many applications beyond research filtering and organization. A selection of supervised methods may be worn for document organization in many domains. It includes large digital libraries of documents, web collections, scientific literature or even social feeds. Hierarchically arranged document collections can be predominantly useful for browsing and retrieval. Opinion mining involves customer reviews or opinions are often short text documents which can be mined to determine useful information from the review. Defined how the classification can be used in order to perform opinion mining is derived. A wide variety of techniques have been designed for text classification used to categorizing the documents.

II. LITERATURE REVIEW

Many types of text representation have been proposed in the past. Information retrieval plays an important role for developing the document search of the adhoc search, filtering, classification [23] and question answers. Many IR models have been developed. There are two major classes in IR history. Global methods and local methods. Global means using corpus based information and local means using set of retrieved or relevant documents. Currently, there are some big research issues in IR and Web search [3], such as evaluation, information needs, effective ranking and relevance. Relevance is a fundamental concept of information retrieval, which is classified into topical relevance and user relevance. The former discusses a document's relevance to a given query; and the latter discusses a document's relevance to a user. Many IR models have been developed for relevance. There are two major classes in IR history: global methods and local methods, where global means using corpus-based information and local means using sets of retrieved or relevant documents. The popular term-based IR models include the Rocchio algorithm, Probabilistic models and Okapi BM25 (more details about Rocchio algorithm and BM25 can be found in Section 6.2), and language models, including model-based methods and relevance models [26]. In a language model, the key elements are the probabilities of word sequences which include both words and phrases (or sentences). They are often approximated by n-gram models [23], such as Unigram, Bigram or Trigram, for considering term dependencies. IR models are the basis of ranking algorithm that is used in search engines to produce the ranked list of documents [6]. A ranking model sorts a set of documents according to their relevance to a give query [23]. For a given query, phrases were very effective and crucial in building good ranking functions with large collections. The data mining techniques are used for text analysis by extracting co occurring terms as descriptive phrases from the document collections. The effectiveness of the text mining systems using phrases as text representation showed no significant improvement. The likely reason was that a phrase based methods had lower consistency of assignment and lower document frequency for terms as mentioned [4]. Pattern mining has been extensively studied in data mining communities for many years. Finding for useful and interesting patterns and rules was still an open problem. Pattern taxonomy model technique was also developed in [11] and [23] to improve the effectiveness by effectively using closed patterns in text mining. A two stage model that used both term based methods and pattern based methods was added [11] in significantly improved the performance of information filtering. Natural language processing is a modern computational technology that can help people to understand the meaning of text documents.

III. PROPOSED WORK

The extracted words from the documents are stored in the feature space. The feature selection involves the indexing tokenizing the text, feature space reduction. There are mainly two approaches in the text categorization knowledge engineering approach and the machine learning approach. In knowledge approach the user defines the rules manually Box of words is one of keyword based method that is widely used. Simplicity is the benefit of this approach. The extracted words from the documents are stored in the feature space. Synonyms and homonyms are the

disadvantage of this approach. The small number of features and over fitting are another issue.

3.1. *Text Document Pre-Processing*

Data preprocessing reduces the size of the input text document significantly. It involves activities like sentence boundary determination, NLP specific stopword removal and stemming [8]. Data preprocessing reduces the size of the input text documents significantly. It involves activities like sentence boundary determination, natural language specific stopword [12] elimination and stemming. Stop words are functional words that can be occur frequently in the language of the text like a, an, the etc. in English language. But this is not useful for classification. Read the whole paper and put all words in the vector. Next again read the file and find contain stopwords then remove similar words from the particular words. Once the data is pre-process it will be the collection of the words that may be in the ontology list. Mining from a preprocessed text is easy as compare to natural languages documents. The preprocessing of documents that are from different sources is an important task text mining process before applying any text mining technique. As text documents are represented as bag of words on which text mining methods are based. Let s be the set of documents & Document = {Word1,word2,….Word n} be the different words from the document set. In order to reduce the dimensionally of the documents words, special methods such as filtering and stemming are applied. Filtering methods remove those words from the set of all words. Stop word filtering is a standard filtering method. Words like prepositions, articles, conjunctions etc. are removed. That contains no informatics used to produce the root from the plural or the verbs. For example Doing, Done, Did may be represented as Do that contain no informatics as such stemming methods: are used to produce the root from the plural or the verbs. For e.g. Doing, Done, Did may be represented as Do. After this method is applied, every word is represented by its root word. Preprocessing text [23] is called tokenization or text normalization. For instance, the following four particular cases have to be considered with care: digits, hyphens, punctuation marks, and the case of the letters. Numbers are usually not good index terms because, without a surrounding context, they are inherently vague. The problem is that numbers by themselves are just too vague. Normally, punctuation marks are removed entirely in the process of lexical analysis. The case of letters is usually not important for the identification of index terms. As a result, the lexical analyzer normally converts all the text to either lower or upper case.

Pre-processing step is crucial in determining the quality of the next stage, that is, the document preprocessing stage. It is important to select the significant keywords that carry the meaning and discard the words that do not contribute to distinguishing between the documents. In the area of text mining, data preprocessing is utilized for extracting, interesting and non-trivial and knowledge from unstructured text data. Information Retrieval (IR) is basically a substance of deciding which documents in a compilation are imaginary to be retrieved and to satisfy the requirement information. The users necessitate intended for information is described through earnings of a query, as well as one otherwise additional search terms, improve an amount of supplementary weight of the sequence words. For this reason, the recovery decision is made by comparing the terms of the query with the index terms, important words otherwise pharses appearing in the document itself.

3.2 Stopwords

The Mutual Information Method (MI)

Stop-word removal is an important preprocessing techniques used in Natural Language processing applications so as to improve the performance of the Information Retrieval System, Text Analytics & Processing System. Stop words are most common words found in any natural language which carries very little or no significant semantic context in a sentence. It just carry syntactic importance which aid in formation of sentence. As a preprocessing operation it must be removed to ease further task and speedup core task in text processing. In order to reduce the dimensionally of the documents words, special methods such as filtering and stemming are applied. Filtering methods remove those words from the set of all words. Stop word filtering is a standard filtering method. Words like prepositions, articles, conjunctions etc. are removed. The mutual information method (MI) is one of the high valuable methods that works by computing the mutual information between a specified expression as well as a document class declared as positive, negative documents. Small common in sequence suggests so as to the expression have a low unfairness authority as well as accordingly it be supposed to be unconcerned.

3.3 Stemming

Krovetz Stemmer (KSTEM)

The Krovetz stemmer was presented in 1993 by Robert Krovetz and is a linguistic lexical validation stemmer. With the intention of it is based on the inflectional possessions of words as well as the language syntax, it is extremely difficult in nature. It successfully as well as precisely replaces inflectional suffixes in three steps:

- Converting the plurals of an expression to extraordinary shape.
- A word can be converting into past tense to present tense.
- Replacing 'ing' from the word like as suffix removal.

The conversion process first removes the suffix and then through the procedure of examination during a vocabulary designed for several recoding and also precedes the stem to a word. The dictionary search for in addition performs several transformations with the intention of to be necessary outstanding to spelling exception as well as in addition converts several stem shaped into a real word, whose significance be capable of to be understood. The power of derivational as well as inflectional examination is in their capability on the way to manufacture morphologically correct stems, suffixes. Stemmer does not discover the stems designed for all statement difference, it is utilized as a pre stemmer before actually applying a stemming algorithm. This would enlarge the momentum as well as usefulness of the most important stemmer method. The Krovetz stemmer is the technique on the way to amplify accuracy in calculation mutually to influence as side to side treating spelling errors as well as worthless stems. Condition the contribution manuscript dimension is great this stemmer becomes weak and does not execute extremely efficiently. The major as well as noticeable mistakes in dictionary based algorithms is their incapability toward deal with by means of words, is not in the lexicon. In addition, a lexicon contain got to be manually shaped in advance that require important efforts. This stemmer does not continually manufacture an expert recall and precision performance.

IV. PATTERN TAXONOMY PROCESS

Pattern can be structured into taxonomy used knowledge discovery model is developed towards applying data mining techniques to practical text mining applications. Knowledge Discovery in Databases (KDD) can be referred to as the term of data mining which aims for discovering interesting patterns or trends from a database. In particular, a process of turning low level data into high-level knowledge is denoted as KDD. The concept of KDD process is the data mining for extracting patterns from data focus on development of knowledge discovery model to effectively use & update discovered patterns and apply it to the field of text mining.

In PTM, split a text into set of paragraphs and exposure every paragraph as a personality transaction, which consists of a position of words. At the succeeding phase, be appropriate the data mining method to discover frequent pattern from these transaction and produce pattern taxonomies. Throughout the pruning phase, non-meaning and redundant prototype are eliminated by applying a proposed pruning scheme. Pattern taxonomy [DIP13] is a tree-like structure that illustrates the relationship between patterns extracted from a text collection. Pattern taxonomy is Text mining utilizes data mining techniques in text sets to discover out connotative knowledge. Its object type is not only structural data other than, also semi structural data or non-structural data. The mining consequences are not simply general situation of one text document but in addition classification and clustering of text sets. The pattern utilized as a word or pharse is extracted as of the text documents. That performs the withdrawal of recurrent sequential patterns. Two parameters are attractive for the method 'SPMining'. The PBPTDM method using different datasets. The most popular utilized data set currently is RCV1, which includes 806,791 news articles for the period between 20 August 1996 and 19 August 1997. These documents were formatted by utilizing a structured XML schema.

The Reuters dataset contains a 1000 unlabeled instances. The Ratio and Random curves are the same. The MaxMin and Simple curves are omitted to ease legibility. The Balanced Random method has a much better precision/recall performance better than the regular Random method, although it is still matched and then outperformed by the active method. For classification accuracy, the Balanced Random method initially has extremely poor performance. The active learning methods had over regular Random sampling were due to this biased sampling. A new querying method called Balanced Random which would randomly sample an equal number of positive and

negative instances from the pool. Obviously in practice the ability to randomly sample an equal number of positive and negative instances without having to label an entire pool of instances first may or may not be reasonable depending upon the domain in question. Random method compared with the Ratio active method and regular random method on the Reuters dataset with a pool of 1000 unlabled instances. TREC filtering track has developed and provided two groups of topics 100 in total for RCV1. The initial group additional 50 topics so as to be collected through human assessors and the subsequent group in addition include 50 topics that were constructed artificially from intersections topics. Every topic alienated documents into two different parts: the training set as well as the testing set. The training set has entirety quantity of 5,129 articles as well as the testing set contains 37,559 articles. Documents within together sets are assigned moreover positive otherwise negative. The "positive" means the document is applicable on the way to the assigned topic. Otherwise "negative" not assigned to the topic. Each and every experimental model utilizes "title" as well as "text" of XML documents only. For dimensionality reduction, stopwords removal is functional as well as is chosen intended for suffix stripping.

V. PBPTDM METHOD

The Proposed method PBPTDM method is used to helps the users to find the huge amount of text documents. The accuracy results have confirmed that all models taking the consideration of the dependency among terms and categories (tf:tcd; pr:tcd) yield the higher accuracy results than others based on document frequency (tf:idf; pr:idf) 77:2% vs. 72:2% and 81:8% vs 73:8%, respectively. It is also possible to conclude the tcd-based methods are more effective than the idf-based methods in text classification. Words may not be the best atomic units, due to one-to-many mappings. Translating words groups helps to resolve ambiguities. It is possible to learn longer and longer pharses based on large training corpora. No need to deal with the complex notions of fertility, insertion and deletions.

K-optimal pattern detection is a data mining method so as to develop another toward the frequent pattern detection approach with the intention of underlies the majority association rule learning techniques. Frequent pattern discovery techniques discover every one pattern proposed for sufficiently recurring in the illustration data. In contrast, k- optimal pattern discovery methods discover the k patterns so as to optimize a user specified calculate of interest. In difference in the direction of k-optimal regulation discovery as well as frequent pattern mining techniques, subgroup discovery focuses on mining interesting patterns with respect to a specified target property of interest. Binary, nominal, or numeric attributes, other than in addition more complex target concepts such as correlations and connecting quite a lot of variables. Background knowledge like constraints and ontological relations can often be successfully applied for focusing and improving the discovery results. Text Mining is the discovery of expensive, so far unknown, information or after the text document. Text classification is the one of the important method to classify the documents to multiple classes. The application of the pattern discovery methods is to identify patterns that characterize a given family of related methods. In this context is need to measure how well distinguish members of the family from non-members based on the occurrence of the pattern. For this purpose a test set consisting of pharse based methods with a well known. Find all occurrences of the motif in the test set and compute the following four scores: TP (true positives) are text document that contain the motif and belong to the family in question, TN (true negatives) are text document that do not belong to the family and do not contain the motif, FP (false positives) are text documents that contain the motif but do not belong to the family and FN (false negatives) are text document that do not contain the motif but belong to the family. Thus TP + TN are the number of correct predictions and FN +FP is the number of wrong predictions.

Based on counts of TP, TN, FP, FN can define various measures. Sensitivity (also called coverage) is defined as TP/(TP+FN) and specificity is defined as TN/(TN+FP). A pattern has maximum sensitivity, if it occurs in all text documents in the relative (regardless of the number of false positives) and it has maximum specificity, if it does not occur in any sequence of other document. Score called correlation coefficient gives overall measure of prediction success. The algorithm SPMining [20] uses the sequential data mining technique with a pruning scheme to find meaningful patterns from text documents. However, it is obviously not a desired method for solving the challenge because of its low capability of dealing with the mined patterns. So that robust and effective pattern deploying technique needs to be implemented.

There are several ways to utilize discovered patterns by using a weighting function to assign a value for each pattern according to its frequency. One strategy has been implemented and evaluated in a pattern mining method that treated each found sequential pattern treat the whole item without breaking them into set of individual terms. Each mined sequential pattern p in PTM.

The following weighting function:

$$W\ p = | da\ da \in D+, p\ in\ da\} || db\ db \in D, p\ in\ db\} |$$

Where da and db denote documents, and D+ indicates positive document in D, such that D+ ⊑ D. However, the problem of this method was the low frequency due to the fact that it is difficult to match patterns in documents especially when the length of the pattern is long. Therefore, a proper pattern deploying method to overcome the low frequency problem is needed

Algorithm for PBPTDM
Step 1: *Taking positive and negative documents to train*
Step 2: *positive document negative document*
Step 3: *for i-▶1 ... n do*
For all I, j, s.t j-i=l do
For all A=X, S do
V pharse [ps] // Pharse the deploying
Step 4: *Sum_supp=0, d< V*
Step 5: *For each pharse pattern p in SP do begin*
Step 6: *Sum_supp+=suppa(p)*
Step 7: *End for*
Step 8: *For each pattern p in SP do begin*
Step 9: *f= suppa(p)/(Sum_supp x len(p))*
Step 10: *V=Sum_supp*
Step 11: *For each term t in p do begin*
Step 12: *P< p U {(t,f)}*
Step 13: *End for*
Step 14: d< d+ p
Step 15: *End*

In order to use semantic information in the pattern taxonomy to improve the performance of closed patterns in text mining to interpret discovered patterns in order to accurately evaluate term weights. The motivation is that discovered patterns that include more semantic meaning than the terms that are selected based on a tem based technique. In term based approaches the evaluation of term weights supports are based on the distribution of terms in documents. The evaluation of term weights is different to the normal term-based approaches. In deploying method, terms are weighted according to their appearances in discovered closed patterns. Terms and global are more likely to gain higher scores than the others. This is due to their high appearance among sequential patterns. However, the patterns support, a useful property of a pattern, is not taken into consideration in pattern deploying method. For instance, the discovered pattern <carbon> acquires an absolute support of 4 in document d1 and 3 in document d4, but the evaluated score for this term is as low as 13/20 compared to 67/60 for another term "emiss" which appears only two more times in supports. Therefore, the support of a pattern is required to be considered while calculating feature significance.

VI. RESULTS AND DISCUSSION

Reshuffle supports of terms within normal forms of d-patterns based on negative documents in the training set. The technique will be useful to reduce the side effects of noisy patterns because of the low frequency problem. This technique is called inner patter evolution here, because it only changes a pattern's term supports with in the pattern. Nevertheless, these PBPTDM method did not yield significant improvements due to the fact that the patterns with high frequency normally the shorter patterns usually have a high value on exhaustivity but a low value on specificity, and thus the specific patterns encounter the low frequency problem. This displays the research on top of the concept of developing an effective Pattern Taxonomy Method toward conquer the aforementioned difficulty through deploying exposed patterns interested in a suggestion liberty. PBPTDM is a pattern based method that depends on the technique of sequential pattern mining as well as utilizes closed patterns because features in the delegate. A noise negative document nd in D_ is a negative document that the system falsely identified as a positive, that is weight (nd)>=Threshold(DP). In order to reduce the noise, need to track which d-patterns have been utilized to give rise to such an error. To reshuffle support of terms within normal forms of discovered patterns based on negative documents in the training set. The technique will be constructive to reduce the side effects of noisy patterns because of the low-frequency problem. This technique is called inner pattern evolution information from the negative has not been exploited during the concept learning there is no doubt that negative documents contains much constructive in sequence to identify ambiguous patterns in the concept.

A set of interesting negative documents, labeled as significant by the system, is first detected. Two types of offenders can be discovered from these interesting negative documents: total conflict and partial conflict.

The basic idea of updating patterns is explained as follow: total conflict offenders are removed from discovered patterns. In support of partial conflict offenders, their term supports are reshuffled within organize toward decrease the belongings of blast documents. The main process of inner pattern evolution is implemented by the IPEvolving. The improvement of IPE is with the intention of all sequential patterns are essential to be concerned for the duration of the developing procedure. The intention of addition establish in the negative documents require on the way to be re-evaluated. The efficiency of the system can be improved. The necessary suggestion of updating patterns is described like: inclusive conflict offenders are unconcerned beginning d-patterns primarily. For fractional conflict offenders, expression supports are reshuffled to organize toward decrease the belongings of blast documents. The main process of inner pattern evolution is implemented by the algorithm IPEvolving. The inputs of this algorithm are a set of discovered patterns DP, a training set D = D+ U D-. The output is a composed of discovered pattern. The second step in IPEvolving is utilized to estimate the threshold for Recall = true positives / (true positives +false negatives)

F-score = 2 * Precision * Recall / (Precision + Recall)

In above true positive means that submit positive document is identify as positive document and false negative means submit positive document is identify negative document and vice versa. False Positive means submit negative document is identifying as positive. In fig.5 explore the Inner pattern evolution. It is used to Shuffling the document. The result of the document after shuffling whether the document is related and unrelated documents. In IPE helps the document using a computer has access to purely random numbers, it is capable of generating a "perfect shuffle", a random permutation of the cards; beware that this terminology (an algorithm that perfectly randomizes the deck) differs from "a perfectly executed single shuffle", notably a perfectly interleaving faro shuffle. From the table it is seen that accuracy for document finding by using pattern mining with the help of keywords gives an effective results. The value of precision and recall F-measure methods used to analyzing the Research papers and Articles. The accuracy value is increased as well as the execution time is reduced.

Table.1 Performance Evaluation of PBPTDM for Single Document

Metrics/Methods	MAP	IAP	Min_Sup
PTM	0.19	0.13	0.13
FPM	0.10	0.14	0.18
PBPTDM	0.19	0.15	0.20

The Pattern Taxonomy Discovery method used to mining the technique with a pruning scheme to find meaningful patterns from text documents. However, it is obviously not a desired method for solving the challenge because of its low capability of dealing with the mined patterns. So that robust and effective pattern deploying technique needs to be implemented

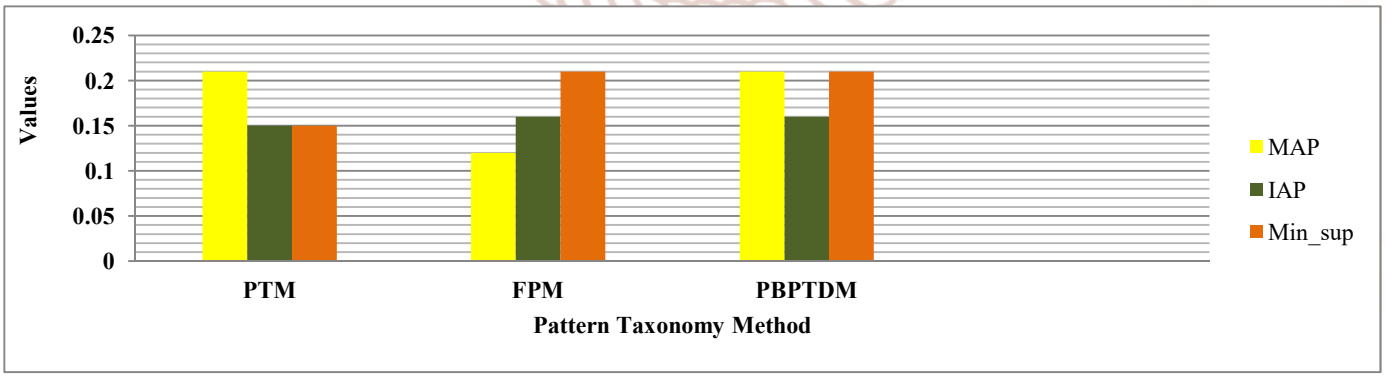

Fig.1 Performance Evaluation of PBPTDM for Single Document

There are several ways to utilize discovered patterns by using a weighting function to assign a value for each pattern according to its frequency. In Fig.1 describes the comparison of Pattern Taxonomy Deploying Model

using for RCV1 data set with five popular categories. Experiments have been conducted on a Pattern taxonomy method, term based pattern taxonomy method, PBPTDM support sets are to evaluate the metrics including MAP, IAP, Minimum-support and resultant datasets are for document accuracy and the values in table 2 compared the pattern taxonomy method utilizing the various types of topics can be analyzed from the RCV1 dataset can be calculated utilizing precision recall and f-measure values. Calculate the frequency measure values compared from various taxonomy methods. The Recall values are 66, 67,65,61,65.4

Table. 2 The Comparison PBPTDM Method for RCV1 Data set

Topic	Precision	Recall	F-Measure	Accuracy
Research article	65	66	73	97.11
Acq	68	67	77	97.23
Wheat	68	65	76	97.45
Earn	67	61	75	97.88
Money-fx	69	65.4	78	98.53

The proposed method is efficient and extracting more sailent features at each scale in the text document such as stemmed, stopword process and pattern discovery methods can be included in the processed. In PBPTDM Shorter pharses occur frequently and are more often applicable to unseen sentences. Longer pharses capture more local context and can be used to translate large chunks of text at one time.

CONCLUSION

Text mining is the process of seeking or extracting the useful information from the textual data. It tries to find interesting patterns from large databases. Text mining is the process of seeking or extracting the useful information from the textual data. It tries to find interesting patterns from large databases. It uses different pre-processing techniques likes stop words elimination and stemming. This paper has given complete information about the text mining preprocessing techniques stop words elimination and stemming algorithms. The discovered knowledge in the field of text mining is having difficulties and ineffective. The reasons are that some useful long patterns with high specificity lack in support. Argue that not all frequent short patterns are useful. The misinterpretations of patterns lead to the ineffective performance so researcher's works for an effective pattern discovery technique has been proposed to overcome low frequency and misinterpretation problems for text mining. The proposed technique uses new model pattern Taxonomy Deploying method to refine the discovered patterns in text documents. The proposed method is efficient and extracting more sailent features at each scale in the text document such as stemmed, stopwords process and pattern discovery methods are included. In PBPTDM phases can occur frequently and are more often applicable to unseen sentences. Longer pharses capture more local context and used to translate large chunks of text at one time.

REFERENCES

1. A Freitas "Comprehensible classification models: a position paper", ACM-SIGKDD Explorations, Pp 27-35, 2014.
2. B.Saini, V.Singh, S.Kumar, "Information Retrieval Models And Searching Methodologies: Survey", In International Journal of Advance Foundation and Research in Computer, Pp 57-62, 2014.
3. Bruce Croft and D. Metzler and T. Strohman. Search Engines: Information Retrieval in Practice. Addison-Wesley, 2009.
4. Been Kim, Cynthia Rudin and Julie Shah, "The Bayesian Case Model: A Generative Approach for Case Based Reasoning and Prototype Classification", Pp 467-498, 2014.
5. B.Ustun and C.Rudin, "Methods and models for interpretable linear classification", Pp 49-61, 2014.
6. B.Narendra, T.Kavitha, P.Surya Chandra, T.BalaKrishna, "A Self Learning Naive Bayes Multi Label Classifier for Analyzing Student Educational Interest", Vol.5 Issue 6, Pp 128-190, June 2015.
7. Benjamin Letham Cynthia Rudin Katherine A.Heller, "Growing a List", Pp 32-97, June 2013.
8. C.Ramasubramanian and R.Ramya, "Effective Pre-Processing Activities in Text Mining using Improved Porter"s Stemming Algorithm", *International Journal of Advanced Research in*

*Computer and Communication Engi*neering, Vol.2, Issue 12, December 2013.

9. Cynthia Rudin, Benjamin Letham, David Madigan, "Learning Theory Analysis for Association Rules and Sequential Event Prediction", Journal of Machine Learning Research , Pp 3385-3436, November 2013.
10. Cynthia Rudin, Benjamin Letham, Ansaf Salleb Aouissi, Eugene Kogan, David Madigan, "Sequential Event Prediction with Association Rules", Pp 3441-3492, 2013.
11. D.S.Guru, Y.H.Sharath, S.Manjunath, "Texture Features and KNN in Classification of Flower Images", IJCA Special Issue on "Recent Trends in Image Processing and Pattern Recognition", Pp 34-47, 2010.
12. H.Dong, F.K.Husain, E.Chang, "A Survey in Traditional Information Retrieval Models", IEEE International Conference on Digital Ecosystems and Technologies, Pp 397-402, 2008.
13. Hassan Saif, Miriam Fernandez,Yulan He, Harith Alani , On Stopwords, Filtering and Data Sparsity for Sentiment Analysis of Twitter.
14. J.Han, M.Kamber, "Data Mining:Concepts and Techniques," Elsevier, Second Edition, Pp 18-25, 2006.
15. J. Xu and W. B. Croft. Improving the Effectiveness of Information Retrieval with Local Context Analysisn. ACM Transactions on Information Systemsl, 18(1):79–112, 2000.
16. Mr.Rahul Patel, Mr.Gaurav Sharma, "A Survey on Text Mining Techniques", International Journal Of Engineering And Computer Science ISSN 2319-7242, Vol.3, Issue 5,Pp 5621-5625,May 2014.
17. Mansi Goyal, Ankita Sharma "An Efficient Malicious Email Detection Using Multi Naive Bayes Classifier", Vol.5, Issue5, Pp39-58, May 2015.