Relevance Pattern Discovery For Text Classification Using Taxonomy Methods

Dr.S.Brindha[1], Dr.S.Sukumaran[2].
[1]Assistant Professor, Dept of Computer Science
[2]Associate Professor, Dept of Computer Science
[1]Palanisamy College of Arts, Perundurai, Tamilnadu, India)
[2]Erode Arts and Science College, Erode,Tamilnadu, India)

Abstract- *Text mining is an instrumental technology that big organizations can employ to extract information and further evolve and create valuable knowledge for more effective knowledge management. It is also an important tool in the web development and text document classification. Pattern Taxonomy learning is an important task for knowledge taxonomy learning task for knowledge acquisition, distribution and classification as well as application expansion and consumption in diverse domains. To reduce the human effort to build a taxonomy from scratch and improve the quality of the learned taxonomy. Pattern Taxonomies are the key to developing successful application domain, such as information retrieval, knowledge searching and classification. In a given domain, the goal of taxonomy learning is to automatically or semi-automatically build pattern taxonomy by identifying domain-specific concepts. In traditional text classification, a classifier is built using labeled training documents of each and every class. This paper studies a problem for negative and positive documents from class p and also other types of documents. The key feature of this problem is that there is no negative document which makes traditional text classification methods is applicable. In this paper, we propose an effective technique to solve the problem.*

Keywords- Classification, Pattern Mining, Pattern Taxonomy, Term Frequency, Relevance Pattern Discovery.

I. INTRODUCTION

Over the past decade, web users have witnessed an exponential growth in the number of web documents are accessible through popular text mining methods. Organizing the large volume of web information in a well-ordered and accurate way is critical for using it as an information resource. One way of accomplishing this in a meaningful way requires document classification. Web document and journal page classification addresses the problem of assigning predefined categories to the web pages by the way of supervised learning. This inaugural learning process automatically builds a model over a set of previously classified web document. The learned model is then used to classify the new web pages. There a large number of classifiers proposed and used for machine learning can be applied for web page classification. These include Support Vector Machine (SVMs), K-Nearest Neighbor (K-NN) and Naive Bayes classifiers. Empirical evaluations of these algorithms on selected small segments of popular web documents. The effectiveness of these algorithms on very large text document and critical journal methods are need to be improve the analyzing the process.

II. RELATED WORKS

The literature consists of successful work done for the range of mining techniques available which are as follows. In the beginning Information Retrieval (IR) provide many keyword-based approach to solve the above challenge, such as probabilistic models[7], Rough set model[9], and Support Vector Machine based filtering model. Scott and Martin [11] also suggested that simple phrase based representations are not worth to pursue, since they found no significant performance improvement. A variety of efficient algorithms such as a priori-based algorithms [3] [9] prefixspan [7], [2], FP-tree [02],[8], SPADE [4], SLPMiner [6], Closet [1] and GST [15] have been proposed. Searching for useful and interesting patterns and discovered patterns is still an open research issue. For this challenging issue, effective pattern based method [3] has been proposed which adopt the concept of closed patterns in text mining.

III. THEORY DISCUSSION

Based on the above Earlier term based methods are provided by Information Retrieval (IR) techniques. The term based methods are classified into rough set models [1], SVM based models [2] and probability models [3]. All the term based methods go through from problems such as synonymy and polysemy. When multiple words have similar meaning, it is called as synonymy. Thus the discovered patterns with term based techniques have semantic meaning and answering the exact user’s requirement is difficult.

Conversely, the experiments in the field of data mining enclose not been proved. Pattern Taxonomy Models (PTMs) came into continuation to defeat the drawbacks of phrase-based mining approaches. Pattern based approaches became replacement other than much improvements are not made to make them more effective for text mining. There are two issues in the regard of effectiveness. They are low frequency and misinterpretation. When terms or patterns are misinterpreted, the result determination not is dependable. Low frequency can't have required support.

Despite the fact that there are some drawbacks, the sequential patterns became promising alternatives to phrases. The reason for this that sequential patterns avail required statistics like terms. Pattern based approaches have some alternatives but much developments are not made to make them more effective for text mining. Over the last many years Information Retrieval (IR) is also utilized to contain several techniques that utilized features of text documents. They are utilized to retrieve content from huge amount of documents based on the terms and their weights. The terms may include different weights based on the context as well. Present might be semantic meanings that are to be considered in IR. Therefore it is not sufficient to only consider weights of terms for document analysis or evaluation. In this paper execute a narrative pattern discovery technique proposed by Zhong et.al.[16]. It first computes specificities of the discovered pattern and then evaluates the weights of terms based on the distribution. Thus it is proficient of neglecting misinterpretation problem. Negative training examples influence is also considered by this in order to avoid low frequency problem. Moreover the ambiguous patterns are update. This phenomenon is known as pattern evaluation. Thus the proposed approach improves accuracy of the discovered patterns.

The planned model is called the relevance feature discovery model, and consists of three major steps: Preprocessing the document and deploying, term and text classification and text weighting. It first finds positive and negative patterns and terms in the training set. Finally, it works out the term weights by using AlgorithmF1 Measure.

IV. PROPOSED METHODOLOGY

Textual documents are gradually more added to the World Wide Web and in addition the electronic databases of organizations. Bag of words approach that makes use of keywords. Tf*idf weighting method used for representing text [6]. In weighting entropy [9] and global IDF are used for text representation in addition to DFIDF. The approach bag of words various schemes were developed for weighting text. The main drawback of this method is that choosing limited number of words is a problem. In a model known as Pattern Taxonomy model is proposed in order to improve the discovered patterns in text mining. In [6] a two stage model was developed. Concept based model came into existence.

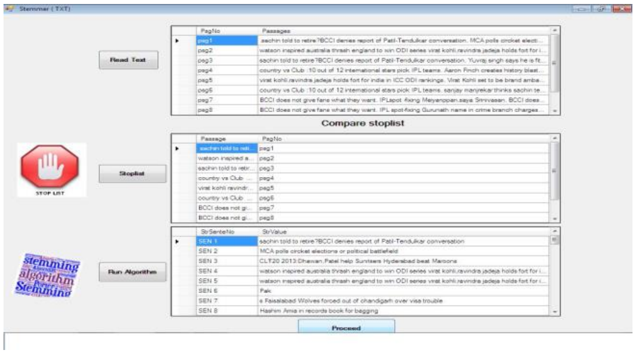
DOCUMENT PREPROCESSING METHODS

A. STOP WORDS REMOVAL

Most of the words are not instructive and that inappropriate are remove from the document representation. (e.g) the a, an, and, there, their, is, was, were, where etc. It is used to develop the competence and possible troubles of removing stop words.

B. STEMMING

Reducing words from their root form. A certain document many contain several occurrences of words like fist, fishes and fishers. An advantage of stemming is to improve the effectiveness to match related words. It is used to reducing the size of indexing to combining words with same roots may reduce indexing size as much as 40-50%. Here utilized portar algorithm is used to stem the words.



	PagNo	Passages
▸	pag1	sachin told to retire?BCCI denies report of Patil-Tendulkar conversation. MCA polls cricket electi...
	pag2	watson inspired australia thrash england to win ODI series virat kohli,ravindra jadeja holds fort for i...
	pag3	sachin told to retire?BCCI denies report of Patil-Tendulkar conversation. Yuvraj singh says he is fit...
	pag4	country vs Club :10 out of 12 international stars pick IPL teams. Aaron Finch creates history blast...
	pag5	virat kohli,ravindra jadeja holds fort for india in ICC ODI rankings. Virat Kohli set to be brand amba...
	pag6	country vs Club :10 out of 12 international stars pick IPL teams. sanjay manjrekar thinks sachin te...
	pag7	BCCI does not give fans what they want. IPLspot fixing Meiyappan,says Srinivasan. BCCI does...
	pag8	BCCI does not give fans what they want. IPL spot-fixing Gurunath name in crime branch charges...

Compare stoplist

	Passage	PagNo
▸	sachin told to reti...	pag1
	watson inspired a...	pag2
	sachin told to retir...	pag3
	country vs Club ...	pag4
	virat kohli ravindr...	pag5
	country vs Club ...	pag6
	BCCI does not gi...	pag7
	BCCI does not gi...	pag8

	StrSenteNo	StrValue
▸	SEN 1	sachin told to retire?BCCI denies report of Patil-Tendulkar conversation
	SEN 2	MCA polls cricket elections or political battlefield
	SEN 3	CLT20 2013:Dhawan,Patel help Sunrisers Hyderabad beat Maroons
	SEN 4	watson inspired australia thrash england to win ODI series virat kohli,ravindra jadeja holds fort for i...
	SEN 5	watson inspired australia thrash england to win ODI series virat kohli,ravindra jadeja holds fort for i...
	SEN 6	Pak
	SEN 7	s Faisalabad Wolves forced out of chandigarh over visa trouble
	SEN 8	Hashim Amla in records book for bagging

Fig.1 Functionality of Stemmer Algorithm

PORTER ALGORITHM:

Step 1 **:** Gets rid of plurals and –ed or –ing suffixes.
Step2 **:** Turns terminal y to i when there is another vowels in the stem.
Step 3 **:** Maps double suffixes to single ones: like –ization, -ational etc.
Step 4 **:** Deals with suffixes, -full, -ness etc.
Step 5 **:** Takes off –ant, -ence, etc.
Step 6 **:** Removes a final -e

The extracted documents from text are converted into Boolean weighting by using the indexing technique of Term Frequency-Inverse Document Frequency. TF-IDF is the product of two statistics, term frequency and converse document frequency. The number of times that terms t occurs in document d. The raw frequency of t by f(t,d) then the simple tf scheme is tf(t,d)=f(t,d). Some other possibilities are including:

- Boolean frequencies: tf(t,d)=1 if t occurs in d and O otherwise
- Logarithmicallyscaled frequency:tf(t,d)=log(f(t,d)+1);
- Augmented frequency, to prevent a bias towards longer documents.

Tf(t,d) =0.5+0.5 Xft (t,d) max {fw,d:w€d}

The inverse document frequency is a calculate of whether the term is familiar or uncommon diagonally all documents. It is obtained by dividing the total number of documents by the number of documents containing the term, and then taking the logarithm of that quotient.*Id(t,D)= log|D||d€D:t€d*||D|:cardinality of D, or the total number of documents in the corpus.|{€D:t€d}|: number of documents where the term t appears. This is common to adjust the formula to 1+|{t€d}|. Mathematically the base of the log function does not matter and constitutes a constant multiplicative factor towards the taken as a whole result. Then TF-IDF is calculated as

tf-idf(t,d,D)= tf(t,d)xidff(t,D).

Our frequency method improved with the weighting assigned to the rare terms. W(t_k,d_j,c_i)=(1-bal).tfidfk,j +bal.tfidfk,j x weighting(2) bal is called a balance factor, which lies between, 0< equal to bal < equal to 1. When bal=0, equation becomes classic TFIDF approach, and when bal=1, equation becomes highly improved approach. Balancing factor is used to get classification taxonomy results. The weighting is the class information and also weighted information to the rare terms. The weighting is calculated using, Weighting=CIx. Here Ai indicates the number of documents belonging to category c_i where the term t_k occurs at least once and C_i denotes the number of documents belonging to category ci where the term t_k does not occur at least once. Fig.2 describes the effective discovery, deploying and evolving of patterns in text mining is used to define the documents into a set of paragraphs like dp1,dp2,…dpn.

C. PATTERN TAXONOMY

A new index structure which allows dealing with a large collection of closed sequential patterns. The direct mining technique, called Pattern Taxonomy Mining (PTM), for building this index during the mining process. The huge amount of closed patterns may hinder to find useful knowledge for the user feedback due to the large search space. A novel indexing mechanism for closed sequential patterns called pattern taxonomy. Pattern taxonomy is a tree like structure that contains a collection of closed sequential patterns and their relations in a feedback set of documents. The main objective of the use of pattern taxonomy is to prepare numerous discovered closed patterns in the feedback set for efficient processing of pattern cleaning.

Pattern Taxonomy Algorithm

```
Train
Step1:  Taking positive(Technology related document) and
        negative(Science related document) documents to
        train
Step2:
Begin
b1= positive document
b2= negative document
if(b1== Tech) //* Tech = Technology document
print("click on positive document to choose");
//select folder for positive document to train
else
if(b2== Sci) //* Sci = Science document
print("click on negative document to choose");
//select folder for negative document to train
Step3:  Now train both the documents
Step4:
if(b1== b2)
print("Select different path to choose positive and
negative
documents");
else
perform Pattern Taxonomy Method to find out
 frequent pattern and weights of the individual
 patterns as well as terms
Test
Step5:  Testing accuracy among the documents which you
         choose
Step6:  Select both documents to compare accuracy
if(b1==b2)
//Result is false positive since you have selected
positive and negative documents for inputs
else
//result is true positive and result is accurate
Prediction
Step7:  Take a document as input
if
//Taken file is related to technology it classifies the
document as positive document
        else
//It is negative document
End
```

Pattern Taxonomy provides a nice way to view possible document representation methods. The new representation captures more important semantics information with closed sequential patterns in text documents and goes beyond the classical term-based representation. Closed sequential patterns have been display to be useful as phrases for information retrieval and text mining.

The concept of pattern taxonomy is employed for document and topic representation in a feedback set. A feedback set of relevant positive documents and non-relevant negative documents, taxonomy profiles of closed sequential and their relations in the feedback set are extracted. Such profiles contain meaningful relevant and non-relevant information for topic representation. The extraction of pattern taxonomy from the large document collections can be infeasible because of the large number of discovered patterns.

D. RELEVANCE PATTERN FEATURE DISCOVERY MODEL

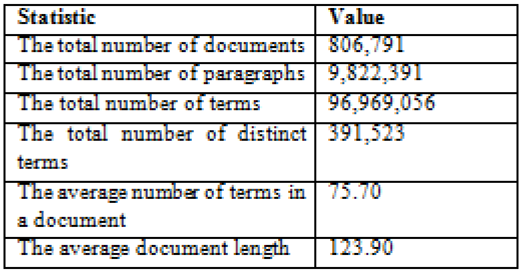
The relevant knowledge is extracted from the feedback set, the next step is utilize the discovered knowledge for the effectiveness of the retrieval system. Information filtering and application in IR and text mining to evaluate the quality of relevant knowledge for a user's topic. In order to use relevance knowledge is to treat patterns as high-level features to find accurate information of user. A new feature weighting method applied to assign accurate weights to each pattern in order to reflect their significance in the user's topic. The relevance of an incoming document is evaluated based on the appearances of the patterns in the document. The representation model for using the relevant knowledge is aimed to address the problem of using specific long patterns in text. We develop equations to deploy high-level patterns over low-level patterns using term support based on their appearance in patterns.

The rational behind this motivation is that discovered patterns include more semantic meaning than the terms that are selected based on a term-based technique (e.g., tf*idf). In term-based approaches, the evaluation of term weights (supports) is based on the allocation of terms in documents. The evaluation of term weights (supports) is different to the normal term-based approaches. PTM is implemented by three main steps: 1) discovering useful patterns by integrating sequential closed pattern mining algorithm and pruning scheme; 2) using discovered pattern deploying; 3) accommodate user profiles by applying pattern evolution.

V. EXPERIMENTAL RESULTS

To evaluate the performance of the proposed PTM model, we added PTM to perform IF tasks and examine the results against those of other methods. The system aims to filter out the non-relevant incoming documents according to the user profiles and extracts profiles from a training datasets. The most frequently used collection for experiments in text categorization and filtering area is the Reuters dataset. Over the years, several versions of Reuters corpora, such as Reuters-21578 and 20 Newsgroups collections has been released. Among the common data collections, Reuters Corpus Volume1 has been most commonly used dataset for the experimental evaluation.

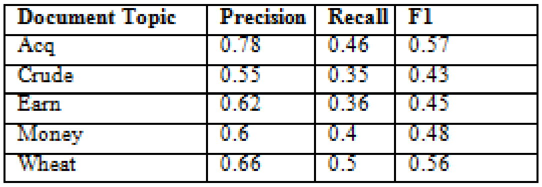
Table.2 The key Statistics of RCV1 Data Collection



Statistic	Value
The total number of documents	806,791
The total number of paragraphs	9,822,391
The total number of terms	96,969,056
The total number of distinct terms	391,523
The average number of terms in a document	75.70
The average document length	123.90

In our experiments, the popular F1 score on the positive class as the evaluation measure. F1 score takes into account of both recall and precision. Precision, recall and F1 defined as,

Table.3 Performance Evaluation



Document Topic	Precision	Recall	F1
Acq	0.78	0.46	0.57
Crude	0.55	0.35	0.43
Earn	0.62	0.36	0.45
Money	0.6	0.4	0.48
Wheat	0.66	0.5	0.56

$$Precision = \frac{\#\ of\ relevant\ terms}{\#\ of\ retrieved\ terms}$$

$$Recall = \frac{\#\ ofrelevantterms}{\#\ oftermscomputed}$$

$$F1 = \frac{2XPrecisionXRecall}{Precision+Recall}$$

For evaluating performance average across categories, macro-average is used. Macro-averaged performance scores are determined by first computing the performance measures per category and then averaging those to compute the global means. Use macro-averaging to find the result of the system.

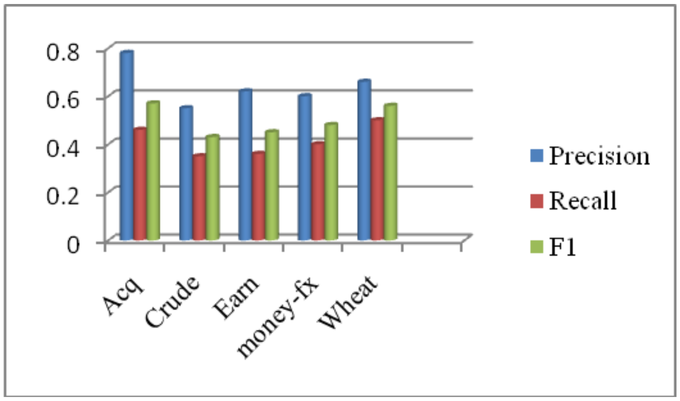


Fig.2 Performance Evaluation Report

Macro average F1 Score is 0.50. This measure is used to know the system's overall performance across sets of data. Compare the effectiveness of relevant patterns obtained by our proposed method. patterns may be easily missed with inappropriate threshold value. The informative patterns by carefully eliminating ambiguous ones. The relevant patterns obtained by the baselines are solely selected from a large pool of positive patterns in relevant documents.

VI. CONCLUSION

Data and Text mining have been around for very long time. The methods are used to discover knowledge based mining like sequential pattern mining, frequent item set mining, closed sequential pattern mining, frequent item set mining, closed pattern mining and maximum pattern mining. But these methods not highly utilized in text mining. For text classification has proposed methods to create virtual examples on the assumption that the label of a document is unchanged even if a small number of words are added or deleted. The experimental results shown that our proposed methods to improve the performance of text classification with Support Vector Machine, especially for small training set. The proposed methods are not readily applicable to NLP tasks other than text classification has been very little studied in NLP, is empirically evaluated.

REFERENCES

[1] Agrawal R, Srikant R.“Fast algorithms for mining association rules in large databases”. In: Proceedings of the VLDB 1994. Morgan Kaufmann; Pp:487-499, 1994.

[2] AggarwalCC,LiY,Wang J, Wang J. “Frequent Pattern mining with uncertain data”. In: Proceedings of the ACM KDD 2009. ACM;Pp:29-37, 2009.

[3] B. S. Everitt, S. Landau, and M. Leese, ”Cluster Analysis”, 4th ed.Wiley Publishing, 2009.

[4] Calders,T, Garboni C, Goethals B. “Efficient pattern mining of uncertain data with sampling”. In: Proceedings of the PAKDD, Part I, Springer; Pp:480-487,2010.

[5] G. Ifrim, G. Bakir, and G. Weikum, “Fast logistic regression for text categorization with variable-length n-grams”, in Proc. ACM SIGKDD Knowl. Discovery Data Mining, Pp: 354–362, 2008.

[6] Heui Lim, “Improving kNN Based Text Classification with Well Estimated Parameters,LNCS”, Vol. 3316, Pp:516-523, oct 2004.

[7] I.Guying and A.Elisseeff, “An Introduction to variable and feature selection”, in J.Mach.Learn.Res., vol. 3, no.1, Pp:1157-1182,2003.

[8] J. Pai, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M. Hsu. “Mining sequential patterns by pattern growth: PrefixSpan Approach”. IEEE Transaction on Knowledge and Data Engineering, vol. 16, no.11, Pp. 1424-1440, 2004.

[9] J. Han, J. Pei, and Y, Yin. “Mining frequent patterns without candidate generation”, in Proc. 2000 ACM-SIGMOD Intl Conf. Management of Data (SIGMOD‟00), Dallas, TX, Pp:1-12, May 2000.

[10] J. Han and K-C, C. Chang, “Data Mining for Web Intelligence” Computer, vol. 35, no. 11, Pp: 64-70, Nov. 2002.

AUTHORS

S.Brindha received B.Sc degree in Science from Bharathiyar University. She done her Master Degree in Information Science and Management in Periyar University and she awarded M.Phil Computer Science from the Bharathiyar University. She has 4 years of teaching experience and 5 years of Technical Experience in Hash Prompt Softwares Pvt. Ltd. She has 2 years of teaching experience to Assistant Professor. At present she is working as Assistant Professor of Computer Science in Palanisamy College of Arts, Perundurai, Tamilnadu, India. She published around 9 research papers in international journals and conferences Her Research area includes Data Mining, Text Mining and Pattern Taxonomy Mining.

Dr. S. Sukumaran graduated in 1985 with a degree in Science. He obtained his Master Degree in Science and M.Phil in Computer Science from the Bharathiar University. He received the Ph.D degree in Computer Science from the Bharathiar University. He has 28 years of teaching experience starting from Lecturer to Associate Professor. At present he is working as Associate Professor of Computer Science in Erode Arts and Science College, Erode, Tamilnadu, India. He has guided for more than 55 M.Phil and 10 research Scholars in various fields. Currently he is Guiding 4 M.Phil Scholars and 6 Ph.D Scholars. He is a member of Board studies of various Autonomous Colleges and Universities. He published around 69 research papers in national and international journals and conferences. His current research interests include Image processing and Data Mining.