Enhanced Closed Sequential Pattern Discovery for Text Mining

Prabha Selvaraj
CSE, Malla Reddy Institute of Engineering and technology, Hyderabad
prabha.dw@gmail.com

***Abstract*—**Enormous amount of data present in the world leads to quandary in retrieval of the useful information. Various algorithms have been proposed to solve the problem of efficient use and mining of the information based on the keywords, phrases, concept of the information need. Pattern Taxonomy Model, Pattern Deploying Method and Pattern Evolving Methods suffer from the problem of low frequency which misleads the mining of useful information from large databases. To solve these problems, a new approach which uses the synset collection in closed sequential pattern mining with the confidence and length constraints. It mines the frequent closed patterns that contain no super-sequence with the same support and uses the synset collection to reduce the redundancy of synonymically similar pattern. It imposes confidence, length constraints to prune the obtained closed sequences to reduce the irrelevant data for mining the useful patterns in large database. Enhanced pattern taxonomy model improves the efficiency and retrieves the relevant information.

***Keywords*—** **Text Mining, Pattern Mining, Constraint based Mining, Information Retrieval.**

I. INTRODUCTION

Text Mining is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources. A key element is the linking together of the extracted information together to form new facts or new hypotheses to be explored further by more conventional means of experimentation. Text mining is different from what are familiar with in web search. Some of the section text mining includes Natural Language Processing, Information Extraction and Information Retrieval. Natural language processing (NLP) is a field of computer science, artificial intelligence, and linguistics concerned with the interactions between computers and human (natural) languages. Information extraction (IE) is the task of automatically extracting structured information from unstructured and/or semi-structured machine-readable documents. Information retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources. Searches can be based on metadata or on full-text indexing. Some of the information retrieval algorithms are Based on Terms, Based on Phrase, Based on Concept and Based on Patterns (Pattern Mining).

Term or keyword based retrieval method is extracting the document based on the keyword present in the document it causes the problem of polysemy and synonym where polysemy means a word has multiple meanings, and synonymy is multiple words having the same meaning. The semantic meaning of many discovered terms is uncertain for answering what users want.

Over the years, people have often held the hypothesis that phrase-based approaches could perform better than the term based ones, as phrases may carry more "semantics" like information. This hypothesis has not fared too well in the history of IR. Although phrases are less ambiguous and more discriminative than individual terms, the likely reasons for the discouraging performance include: 1) phrases have inferior statistical properties to terms, 2) they have low frequency of occurrence, and 3) there are large numbers of redundant and noisy phrases among them.

Concept based retrieval method allows users to post their query by either simply using keywords or by using a form of natural language. This model included three components. The first component analyzed the semantic structure of sentences; the second component constructed a conceptual ontological graph (COG) to describe the semantic structures; and the last component extracted top concepts based on the first two components to build feature vectors using the standard vector space model. The advantage of the concept-based model is that it can effectively discriminate between non important terms and meaningful terms which describe a sentence meaning but it is very complex to access the document.

Pattern mining concentrates on identifying rules that describe specific patterns within the data. Pattern Mining is used to identify items that were often occurred together and also it identifies items which are frequently occurring. Various types of Pattern Mining are Frequent Pattern Mining, Structured Pattern Mining and Sequential Pattern Mining. Frequent pattern mining is in the form of association rule mining. It analyses document which occurred frequently by finding associations between the different terms that wants to retrieve. Some of the algorithms for frequent pattern mining are: Apriori, FP-Growth, Eclat, CHARM, CLOSET, CLOSET+, FPClose, AFOPT, CARPENTER, COBBLER, TD-Close etc.

Structure mining or structured data mining is the process of finding and extracting useful information from semi structured documents. Graph mining is special cases of structured data mining such sophisticated patterns go beyond sets and sequences, toward trees, lattices, and graphs. As a general data structure, graphs have become increasingly important in modeling sophisticated structures and their interactions, with broad applications including chemical informatics, bioinformatics, computer vision, video indexing, text retrieval, and Web analysis. Some of the algorithms are used for the structured pattern mining algorithms are SUBDUE, AGM, FSG, gSpan, MoFa, SPIN, Gaston, TreeMinerV, FREQT, CloseGraph, CloseCut, Splat, TSMiner, GREW etc,.

Sequential Pattern Mining finds interesting sequential patterns among the large collection of documents. It finds out frequent subsequences as patterns from a sequence documents. Some of the techniques are Apriori-based Method, Pattern-growth Method, Approximate sequential pattern mining, Constrained Based sequential pattern mining, Maximum sequential pattern mining, Incremental sequential pattern mining and Closed sequential pattern mining. Most of them are proposed for the purpose of developing efficient mining algorithms to find particular patterns within a reasonable and acceptable time frame. With a large number of patterns generated by using the data mining approaches, how to effectively exploit these patterns is still an open research issue.

To solve this issue the closed sequential pattern mining approach was introduced. Instead of mining complete set of frequent subsequence's it mines frequent closed subsequences only, i.e., those containing no super-sequence with the same support. Some of the algorithms of closed sequential pattern mining are CLOSPAN, BIDE, Par-CSP, TSP, COBRA and CEMiner. Page Layout.

II. LITERATURE REVIEW

Text mining is the technique that helps users find useful information from a large amount of digital text documents on the Web or databases. Traditional Information Retrieval (IR) has the same goal of automatically retrieving relevant documents as many as possible while filtering out non-relevant ones at the same time. Many text mining methods have been developed in order to achieve the goal of retrieving useful information for users.

Many types of text representations have been proposed in the past. Most of the approaches are based on the two empirical observations regarding text that are more times a word occur in a document or throughout all documents in the collection. Aas and Eikvil proposed many weighting scheme [1] to weight the term such as, Boolean Weighting, Word Frequency Weighting , tf idf – Weighting , tfc –Weighting, ltc-Weighting, Entropy weighting. Using the tfidf weighting scheme and relevance feedback of the query is categorized to the text was evaluated by Joachims [6]..

The disadvantages of the bag of words are not scalable well in the large values of terms and also it does not hold of many sophisticated learning algorithms used for classifier induction. To solve problem of over fitting and scalable in bag of words various indexing and dimensionality reduction measures are used such as, DIA association factor, Information gain, Mutual information, Chi-square, NGL coefficient Latent Semantic Indexing etc was investigated by Lewis and Sebastiani ([7],[15]).

The traditional text classification, a classifier is built using labeled training documents of every class it labels only for the positive documents its leaves other than the positive documents to solve this approach the combination of Rocchio and support vector machine are proposed by Li and Liu [8].In the past a number of IR researchers have expressed their unsatisfaction with the set of words approach, and have tried to use notions of what a feature is that are at the same time semantically richer and technically feasible. It solves the problem by using phrases was proposed by Caropreso, Matwin and Sebastiani([4], [15]), co-occurring terms was projected by Ahonen et.al. [3], and also semantic relationship between terms using RIPPER and WORDNET tools instead of terms was investigated by Scott and Matwin [14].

The phrases are lower frequency than the terms so moving to the techniques terms based ontology mining approaches it uses some of the methods such as hierarchical concept clustering, Dictionary Parsing etc, it order to exploits the similarity of the items into hierarchical cluster of items was projected by Maedche [10]. Even though the advantage of ontology it also having some problems to overcome that phrase based retrieval method was introduced.

Mining of patterns are interesting research topic in information retrieval. Some of the researches introduces various algorithms such as, Apriori like algorithm such as Apriori, AprioriTid, AprioriHybrid was introduced by Agrawal and Srikant [2] that causes the problem of huge candidate generation, finding comparative sentences was introduced by Park , Chen and Yu [12] that causes the problem of low occurrence of patterns, pattern growth algorithms such as FP- Tree was proposed by Han, Pei and Yin [5] that causes the problem of not fit in memory, Projected based algorithms such as freespan and prefixspan was introduced by Pei et.al [13] that cause the problem of waste of memory in projected database, SPADE algorithm it process as , splits the larger database into smaller one and process the operations was demonstrated by Zaki [21] it can suffer the problem of complexity of operation , Clospan it works as mines the closed itemset only i.e. those containing no super sequences with same support was established by Yan , Han and Afshar [20]. SLPMiner it mines the database with the constraints of the length was projected by Seno and Karypis [16] but it suffers the problem setting the constraints.

A novel concept for mining text documents for sequential patterns was introduced by Wu et.al [19]. Instead of using single words, use pattern-based taxonomy to represent documents. By pruning meaningless patterns, which have been proven to be the source of the 'noise' in this study, the problem of 'overfitting' is solved.

Pattern Taxonomy Model is obviously not a desired method for conquering the challenge because of its low capability of dealing with the mined patterns. In order to solve the challenges more robust and effective pattern deploying techniques was proposed by Wu, Li and Xu [18]. In that approach it refines the patterns twice to deploy the discovered patterns into a feature space which is used to represent the concept of documents.

It is a still challenging issue for PTM to deal with low frequency patterns because the measures used from data mining (e.g., "support" and "confidences") to learn the profile turn out be not suitable in the filtering stage. To solve the problem of PTM , Li et.al. [9] introduced the two stage filtering model in that model it consists of two stages : first stage is topic filtering stage to solve the problem of mismatch, and the second stage is pattern taxonomy mining to solve the problem of overload.

Natural Language Processing (NLP) is both a modern computational technology and a method of investigating and evaluating claims about human language itself. Recently, a new concept-based model was presented by Shehata, Karray and Kamel [17] to bridge the gap between NLP and text mining, which analyzed terms on the sentence and document levels.

The concept-based model analyzes each term within a sentence and a document using the three components namely, concept based statistical analyzer to analyzes each term on the sentence and the document levels , Conceptual Ontological Graph is to captures the semantic structure of each term within a sentence and a document, and then concept extractor, combines the two different weights computed by the concept based statistical analyzer and the COG representation to denote the important concepts with respect to the two techniques.

Ning Zhong, Yuefeng Li and Sheng-Tang Wu [11] was proposed two methods such as pattern deploying and inner pattern evolving methods to improve the effectiveness of using and updating discovered patterns, which first calculates discovered specificities of patterns and then evaluates term weights according to the distribution of terms in the discovered patterns rather than the distribution in documents for solving the misinterpretation problem. Compared with other methods, pattern deploying and inner pattern evolving methods finds the relevant and interesting information from the documents. Paragraphs must be indented.

III. ENHANCED PATTERN TAXONOMY MODEL

The system architecture of the proposed model consists of the following functions: User interface, Query Processing System, Pattern Identifier and Clustering and Document Retrieval Manager. In pre-processing stage, this turns the text documents into a keyword-based representation with the help of two processes namely Stop Word Removal and Stemming Methods. The pre-processed result is passed to the Pattern Identifier and Clustering (PIC) process. PIC processes the data and finds the weighted constraint based closed patterns. The Results are clustered by using the pattern clustering methods which groups the patterns based on the interpattern similarity measures and stores it in the database.

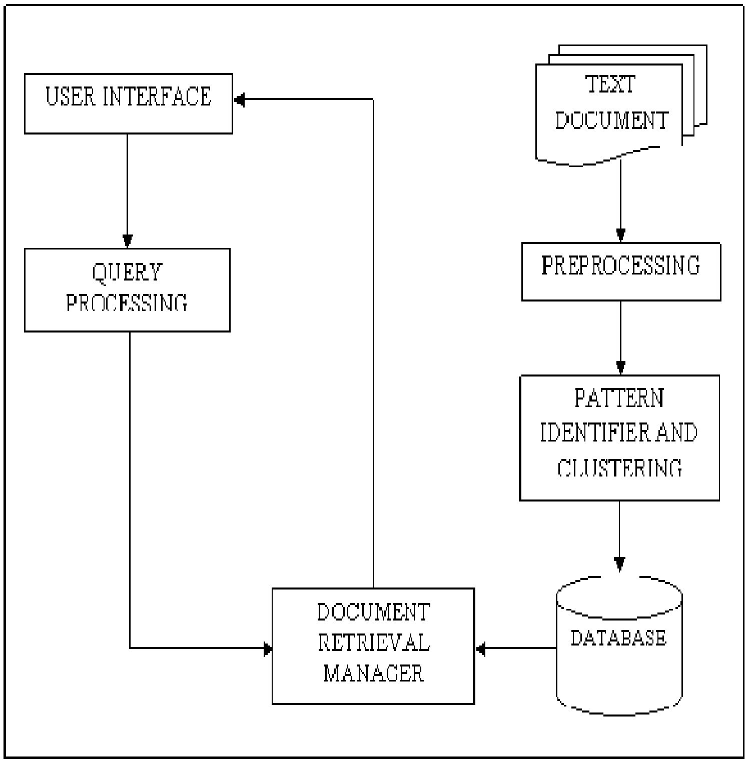


Fig.1 System Architecture

ALGORITHM FOR PREPROCESSING
Algorithm. Pre-processing (D, sw, sl)
Input: D Text Document, sw Stop Words, sl Synset List
Output: pl Pattern list of various combinations

```
Method:
For each Document D
begin
        p= Stopwords(sw,D)
end for
For each Pattern p Є D
begin
        Synset(sl,p)
        Stemming(p)
        count occurrence (p Є D)
        pl= possibleCombination(p)
end for
Algorithm. Stopword(sw,D)
Input: sw Stopwords, D Text Documents
Output: p Patterns
Method:
For each Document D
begin
        if sw matches with  Document  D containing pattern  p
                remove pattern p
        end if
        else
                return p
        end else
end for

Algorithm. Synset(sl,p)
Input: sl Synsetlist Database, p Patterns
Output: p Patterns
Method:
For each pattern p
begin
        if pattern p matches with database sl
                form the similarity tree for the given patterns
        end if
end for

Algorithm. Stemming(p)
Input: p patterns
Output: p patterns
Method:
For each pattern p
begin
        apply porter stemmer steps
end for
```

ALGORITHM FOR PATTERN IDENTIFIER AND CLUSTERING

```
Algorithm.PIC(pl,s,c)
Input: pl Pattern list of various combinations, s Minimum Support, c Minimum Confidence
Output: cd Clustering of Documents, lb label for each cluster
```

```
Method:
if pl $\neq \emptyset$
        /* Pattern Nomenclature Model*/
        pl' = {$\emptyset$} /* Intialize the pattern list pl' as null*/
        Assume length of patterns as 3/*Setting length constraints*/
    for each pattern p in pl
        begin
        generate p-projected database PD /*Generates the patterns in Hierarchical Manner*/
        for each frequent pattern fp in PD
        P' = p $\bowtie$ fp /* Combines the patterns with frequent patterns*/
        if support(P') $\neq \emptyset$ /*Checks the support of patterns are not null*/
        pl' = pl' $\cup$ P' /*Stores the patterns combined patterns*/

end if
        end for
        end for
        /* Support and confidence Calculation*/
        for each frequent pattern P' in pl'
                for each (p,w) in pl' do
                s= support(p) +w
                c= support($p_i$,$p_j$)/support($p_i$)
                end for
        end for
        /* Constraint Applying*/
        for each frequent pattern P' in pl'
                if  pl' $\geq$ (s $\cup$ c)
                        update pl'
                else
                        remove pattern from pl'
                end if
        end for
        /* Clustering the Document*/
        for each pattern p in pl' Є Document D
        finds the similarity measures of patterns for different documents for different documents {D1,D2,…..,Dn}
        if  Similarity(D1,D2…Dn) have maximum value Cluster the documents
        else
        Form a new cluster
        end if
        Label the cluster  (lb) by maximum occurred patterns in the cluster
        end for
        end for
else
        return null
end if
```

ALGORITHM FOR QUERY PROCESSING

Algorithm. Query Processing(q, sw, sl)

Input: q User Query, sw StopWords, sl Synset List

Output: qk Query Keyword

```
Method:
For given user query q
begin
            qk = Stopwords(sw,q)
end for
For each query keyword qk Є q
begin
            Stemming(p)
end for
```

ALGORITHM FOR DOCUMENT RETRIEVAL MANAGER

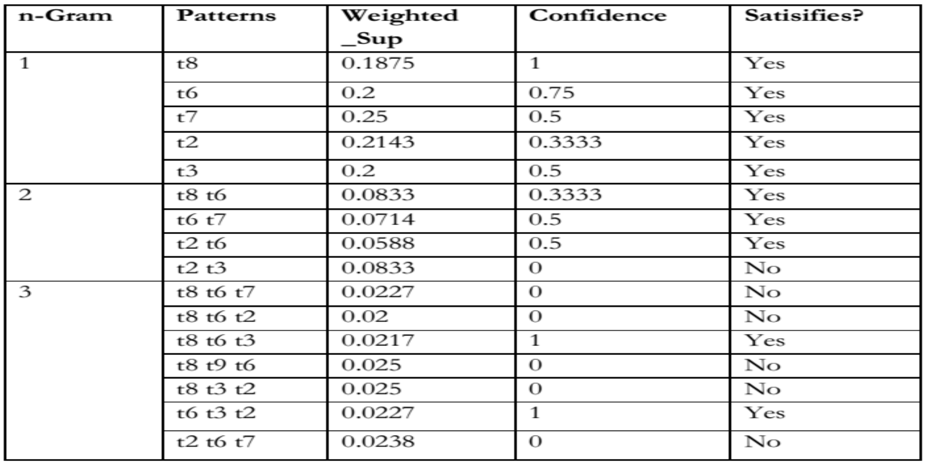
```
Algorithm. DRM(qk, cd, lb)
Input:  qk QueryKeyword, cd Clustered Documents stored in Database, lb label for each cluster
Output: sr Search Results

Method:
For each query keyword qk in q
begin
            if qk matches with lb
            order the cluster by most occurrence of the qk in clustered
            display the whole cluster in order
            else
            display the no records found
            end if
end for
```

ALGORITHM FOR MAXIMUM CAPTURING

```
Algorithm.maxcap (D, p)
Input: D documents, p Patterns
Output: cl Clusters of Documents
Method:
Construct the similarity Matrices for documents with patterns as S
Min_value(S) = Min(S) /* Finds the Document pair containing Minimum value in Matrices S*/
Max_value(S) =Max(S) /* Finds the Document pair containing Maximum Value in Matrices S*/
if  Max_value(S) ≠  Min_value (S) /* Checks the maximum value pair with minimum value pair */
      For  each Max_value(S)
      begin
            if  Max_value(S) ϵ Cluster cl  /*Checks whether the Document pair already present in Existing Cluster or not*/
               cl=cl+Max_value(S)  /* Add the Document pair with Existing Cluster*/
           else
               Form a new cl
           end if
          Set Max_value(S)=0  /* After assigning the Document Pair into Cluster set Max_Value as Zero*/
      end for
end if
if any document D not in any cluster cl
      Form a new cl
end if
```

TABLE 1
N-GRAM WITH MIN_WEIGHTED_SUP=2% AND MIN_CONFIDENCE=2%



n-Gram	Patterns	Weighted _Sup	Confidence	Satisifies?
1	t8	0.1875	1	Yes
	t6	0.2	0.75	Yes
	t7	0.25	0.5	Yes
	t2	0.2143	0.3333	Yes
	t3	0.2	0.5	Yes
2	t8 t6	0.0833	0.3333	Yes
	t6 t7	0.0714	0.5	Yes
	t2 t6	0.0588	0.5	Yes
	t2 t3	0.0833	0	No
3	t8 t6 t7	0.0227	0	No
	t8 t6 t2	0.02	0	No
	t8 t6 t3	0.0217	1	Yes
	t8 t9 t6	0.025	0	No
	t8 t3 t2	0.025	0	No
	t6 t3 t2	0.0227	1	Yes
	t2 t6 t7	0.0238	0	No

A User Interface is the system by which users interact with a system. The user interface includes physical and logical components. User interfaces exist for information retrieval systems, and provide a means of 1) Input, allowing the users to posting the search query. 2) Output, allowing the system to display the resulted search documents. The Query Processing System (QPS) does two function, first is Query Analyzer which analyzes the query given by the user, process the query and get the query terms. The next function is Query word extractor, which extracts the keywords contain in the query terms and sends to the Document Retrieval Manager (DRM). The Document Retrieval Manager consists of three processes: 1) Word Pattern Analysis which analysis the keyword from query processing system and patterns in the database based on the pattern identifier and clustering. 2) Document Retrieval which retrieves the text documents which matches the query keyword and 3) Page Ranking which ranks the document based on the relevancy of the query keywords to the text document. Finally the search results are displayed to the user. The algorithm PIC gives the process of the pattern identifier and clustering module.

IV. RESULT AND DISCUSSION

TABLE 2.
LIST OF METHODS USED FOR EVALUATION

Method	Description	Algorithm
Sequential Patterns	Text Mining using Sequential Patterns	SPM(Sequential Pattern Mining)
Closed Sequential Patterns	Text Mining using Closed Sequential Patterns	CSPM(Closed Sequential Pattern Mining)
n-Gram	n-Gram with n=3	3 Gram
Rocchio	Rocchio Method with α=1 and β=0	$\vec{c}=\alpha\frac{1}{\lvert D^{+}\rvert}\sum_{\vec{d}\in D^{+}}\frac{\vec{d}}{\lVert\vec{d}\rVert}-\beta\frac{1}{\lvert D^{-}\rvert}\sum_{\vec{d}\in D^{-}}\frac{\vec{d}}{\lVert\vec{d}\rVert}$
Probablistic	Probablistic method	$W(t)=\log(\frac{r+0.5}{R-r+0.5}/\frac{n-r+0.5}{(N-n)-(R-r)+0.5})$
TF-IDF	TF-IDF Method	W(t) = TF(d,t) X IDF(t)

TABLE 3.
COMPARISON OF PERFORMANCE MEASURES WITH EPTM

Method	Precision	Recall	$F_{\beta=1}$	b/p	MAP
EPTM	0.89	0.62	0.655	0.775	0.6956
Sequential Patterns	0.57	0.524	0.566	0.66	0.6345
Closed Sequential patterns	0.6	0.543	0.576	0.7	0.6453
nGram	0.57	0.524	0.563	0.66	0.6345
Rocchio	0.63	0.543	0.623	0.705	0.6435
Prob	0.61	0.552	0.582	0.724	0.6623
TF-IDF	0.43	0.5145	0.5423	0.523	0.5235

Measures

Precision is the fraction of the documents retrieved that are relevant to the user's information need.

$$\text{Precision} = \frac{\{\text{relevant documents}\}\cap\{\text{retrieved documents}\}}{|\{\text{retrieved documents}\}|}$$

Recall is the fraction of the documents that are relevant to the query that are successfully retrieved.

$$\text{Recall} = \frac{|\{\text{relevant documents}\}\cap\{\text{retrieved documents}\}}{|\{\text{relevant documents}\}|}$$

The weighted harmonic mean of precision and recall, the traditional F-measure or balanced F-score the general formula for non-negative real β is:

$$F_\beta = \frac{(1+\beta^2).(\text{Precision}.\text{Recall})}{(\beta^2\ \text{Precision}+\text{Recall})}$$

Mean average precision for a set of queries is the mean of the average precision scores for each query.

$$\text{MAP}= \frac{\sum_{q=1}^{Q}\text{AveP}(q)}{Q}$$

Where Q is the number of queries.

Break Even Point (b/p) is when precision equals to recall

TABLE 4.
COMPARISON OF PERFORMANCE MEASURES OF PTM WITH EPTM

Method	EPTM	PTM
Precision	**0.890**	**0.880**
Recall	**0.628**	**0.568**
$F_{\beta=1}$	**0.655**	**0.652**
b/p	**0.775**	**0.756**
MAP	**0.695**	**0.675**

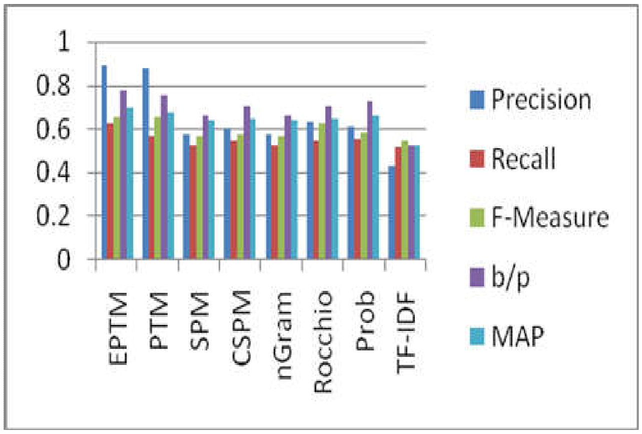


Fig. 2 Comparison of different methods with EPTM

V. CONCLUSION

Information retrieval is difficult due to presence of large amount of data in the world. Various approaches are introduced such as Pattern Taxonomy Model, Pattern Deploying Model etc., to solve the difficulties but it suffers the problem of low frequency as well as mining of large database are difficult. To improve the system, a new method is EPTM is proposed in which it uses closed sequential patterns with the length and confidence as constraints. This approach reduces the irrelevant data and mines the useful information with the help of enhanced closed sequential pattern from the large database.

In Pattern Taxonomy Model, it uses some of the features such as support, confidence, Relationship between patterns, Distribution of pattern taxonomies and the dimensions of taxonomies are evaluated but in this model only few features has been implemented and evaluated.

.

REFERENCES

[1] Aas K and Eikvil L. "Text Categorisation: A Survey", Technical Report Raport NR 941, Norwegian Computing Center, 1999.

[2] Agrawal R and Srikant R. "Fast Algorithms for Mining Association Rules in Large Databases", Proc. 20th Int'l Conf. Very Large Data Bases (VLDB '94), pp. 478-499, 1994.

[3] Ahonen H., Heinonen O., Klemettinen M and Verkamo A.I. "Applying Data Mining Techniques for Descriptive Phrase Extraction in Digital Document Collections", Proc. IEEE Int'l Forum on Research and Technology Advances in Digital Libraries (ADL '98), pp. 2-11, 1998.

[4] Caropreso M.F., Matwin S and Sebastiani F. , "Statistical Phrases in Automated Text Categorization", Technical Report IEI-B4-07-2000, Instituto di Elaborazione dell'Informazione, 2000.

[5] Han J., Pei J and Yin Y. "Mining Frequent Patterns without Candidate Generation", Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD '00), pp. 1-12, 2000.

[6] Joachims T. "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.

[7] Lewis D.D. "Feature Selection and Feature Extraction for Text Categorization", Proc. Workshop Speech and Natural Language, pp. 212-217, 1992.

[8] Li X and Liu B. "Learning to Classify Texts Using Positive and Unlabeled Data", Proc. Int'l Joint Conf. Artificial Intelligence (IJCAI '03), pp. 587-594, 2003.

[9] Li Y., Zhou X., Bruza P., Xu Y and Lau R.Y."A Two-Stage Text Mining Model for Information Filtering", Proc. ACM 17th Conf. Information and Knowledge Management (CIKM '08), pp. 1023-1032, 2008.

[10] Maedche A. "Ontology Learning for the Semantic Web. Kluwer Academic", 2003.

[11] Ning Zhong., Yuefeng Li and Sheng-Tang Wu. "Effective Pattern Discovery for Text Mining", IEEE Trans. Knowledge and Data Eng., vol.24, no.1, 2012.

[12] Park J.S., Chen M.S and Yu P.S. "An Effective Hash-Based Algorithm for Mining Association Rules", Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD '95), pp. 175-186, 1995.

[13] Pei J., Han J., Mortazavi-Asl B., Pinto H., Chen Q., Dayal U and Hsu M. "Prefixspan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth", Proc. 17th Int'l Conf. Data Eng. (ICDE '01), pp. 215-224, 2001.

[14] Scott S and Matwin S. "Feature Engineering for Text Classification", Proc. 16th Int'l Conf. Machine Learning (ICML '99), pp. 379-388, 1999.

[15] Sebastiani F. "Machine Learning in Automated Text Categorization", ACM Computing Surveys, vol. 34, no. 1, pp. 1-47, 2002.

[16] Seno M and Karypis G. "Slpminer: An Algorithm for Finding Frequent Sequential Patterns Using Length-Decreasing Support Constraint", Proc. IEEE Second Int'l Conf. Data Mining (ICDM '02), pp. 418-425, 2002.

[17] Shehata S., Karray F and Kamel M. "A Concept-Based Model for Enhancing Text Categorization", Proc. 13th Int'l Conf. Knowledge Discovery and Data Mining (KDD'07), pp. 629-637, 2007.

[18] Wu S.T., Li Y and Xu Y. "Deploying Approaches for Pattern Refinement in Text Mining", Proc. IEEE Sixth Int'l Conf. Data Mining (ICDM '06), pp. 1157-1161, 2006.

[19] Wu S.T., Li Y., Xu Y., Pham B and Chen P. "Automatic Pattern- Taxonomy Extraction for Web Mining", Proc. IEEE/WIC/ACM Int'l Conf. Web Intelligence (WI'04), pp. 242-248, 2004.

[20] Yan X., Han J and Afshar R. "Clospan: Mining Closed Sequential Patterns in Large Datasets", Proc. SIAM Int'l Conf. Data Mining (SDM '03), pp. 166-177, 2004.

[21] Zaki M. "Spade: An Efficient Algorithm for Mining Frequent Sequences", Machine Learning, vol. 40, pp. 31-60, 2001.