ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 15, March 2016

A COMPARATIVE STUDY ON PATTERN MINING TECHNIQUES

S.Brindha[1], Dr.K.Prabha[2] , Dr.S.Sukumaran[3].

[1](Ph.D Scholar, brindha.balajiee@gmail.com, Department of Computer Science Erode Arts and Science College, Erode, Tamilnadu, India)
[2](Assistant Professor of Computer Science, prabhaeac@gmail.com, Periyar University PG Extension Centre, Dharmapuri, Tamilnadu, India)
[3](Associate Professor, prof_sukumar@yahoo.co.in, Department of Computer Science, Erode Arts and Science College, Erode,Tamilnadu, India)

ABSTRACT: *Text mining is the innovation of exciting facts in text documents. It is a difficult concern to hit upon accurate information in text documents is used to facilitate users to find what the users necessary. Using Text mining the users can able to find and filtered useful information from a large amount of digital data. Due to the continuous growth of the volume of the text data automated of implicit, formerly unknown and potentially useful in sequence becomes more necessary to properly useful this infinite resource of information. Text mining consequently corresponding to the extension of the data mining approach to textual data and is concerned with a variety of tasks, such as removal of the information absolutely contained in collection of document or similarity based on structuring. The collection of text in common lacks the forced structure of an established database. The text can be expressed in a vast range. The text information can be encodes the information in a form that is difficult to decipher automatically. Pattern based methods can give an outstanding performance in keyword based and pattern based methods. The aim of this paper is to compare various pattern mining techniques for analysing text documents, such as frequent item sets, closed sequential pattern, PTM, PDM, IPE.*

Keywords: *Text Mining, Pattern Mining, Pattern Taxonomy Model, Pattern Deploying Method, and Inner Pattern Evoluation.*

I. INTRODUCTION

Mostly text mining methods use keyword based or phrase based techniques. But it has lot of drawbacks. Sequential closed pattern mining technique is used to find the patterns in a particular document. Pattern taxonomy model is used for the purpose of documents can be used effectively using discovered patterns. Pattern taxonomy model reconsider the method of patterns besides the deploying the documents based on their parallel relations in the pattern taxonomies. In addition to the pattern deploying, Text mining is used to finding and to discover the interesting knowledge in text documents. It is a challenging issue to find perfect knowledge in text documents to help users to find what are they want. Many applications, such as market based analysis and business based management, can benefit by the use of the information and the big data are to be extracted from knowledge. Knowledge discovery can be effectively use and update discovered in databases, it means data mining is having all methods of knowledge discovery process and present a modeling phrase that is application of methods and algorithms for computation of investigate pattern or models. In the precedent method, an important and essential number of data mining techniques contain presented. In order to perform different and more meaningful knowledge tasks. The techniques are included in association rule mining, frequent item set mining, in order pattern mining, maximum pattern mining and closed pattern mining. The purpose of developing efficient mining algorithms to find particular patterns within a logical and suitable time frame. With a large number of patterns generated by using the text mining approaches, how to successfully make use of these patterns is still an open research issue. Text mining is parallel to data mining, but it is useful to free semi structured and unstructured texts. It should be capable of classifying documents converted into taxonomy, also trade with the machine supported analysis of text. It uses PTM techniques from information retrieval, information extraction as fit while natural language processing (NLP) and connects them with the algorithms and methods of Databases in knowledge discovery, statistics and also machine learning [2][11]. The Pattern mining approach mostly depending upon the classification concepts. Generally classification is one of the types of Supervised Learning model. This can be used to extract models relating necessary data classes or envisage potential trends. This model is used to predict an objective value or aspect. It is also describe the background knowledge. Challenging issue of using phrases: Finding useful phrases for text mining and classification [24], since naturally, phrases have inferior statistical properties of words, and there are a big number of redundant and noisy phrases. Second challenge is patterns can be a promising alternative phrases. Like words, patterns treat like good statistical properties. Text mining has developed some techniques: maximal patterns, closed patterns and master patterns for removing out of work and noisy patterns.

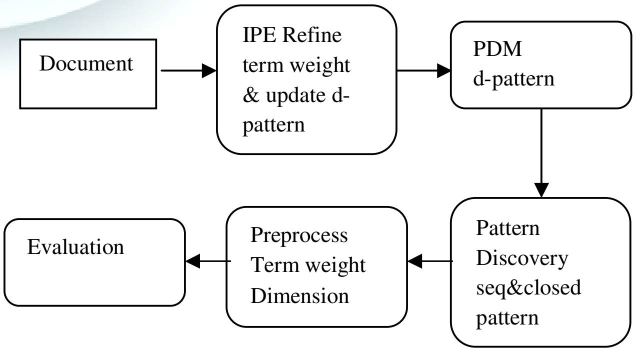
Document
IPE Refine term weight & update d-pattern
PDM d-pattern
Pattern Discovery seq&closed pattern
Preprocess Term weight Dimension
Evaluation

Fig.1.Pattern Mining Method

Pattern mining method includes modules like pattern discovery, pattern deployment, and pattern evaluation. Fig1. shows process from the input documents.The text document is XML format. Documents in both the sets are positive or negative. Positive means document is related to the topic otherwise negative. By using this method sequential closed frequent and non sequential closed frequent patterns from the dataset. The document is divided into paragraph. These are processed for pre-processing step.

With the minimum support dimension reduction process is done to get feature set. Sequential pattern and closed sequential pattern is discovered by pattern discovery model. Pattern discovery evaluates specificities of pattern and then evaluates term weights according to the distribution. It improves the accuracy of term weights the discovered pattern are more special than whole process documents. The pattern discovery pattern deploying method summarizes discovered pattern using PDM. The d-pattern is used to discover all pattern in positive documents are composed. The inner pattern evolution is to identify the noisy patterns in documents and update D-pattern by shuffling. On the basis of refined pattern weights concept formation is done.

II. PATTERN TAXONOMY MODEL (PTM)

Patterns can be structured into a taxonomy using the “is-a” relation method. Pattern Taxonomy [13] based model of the containing sequential pattern used to perform the task. Instead of the keyword-based concept used in the established document representation model, the pattern based model contains frequent sequential patterns are use to perform the related concept of task. In PTM, split a text into set of paragraphs and treat each paragraph as an individual transaction it considered to be a set of words. Each and every paragraph consists of a set of words and is used as an individual transaction. And to select extract patterns from the transactions at the subsequent phrase. The data mining and text mining method is applied to each paragraph to extract the particular patterns. All the documents can be consists of positive (D+) and negative (D-) documents. At the subsequent phrase, apply the data mining method to find frequent pattern from these transaction and generate pattern taxonomies. There are two main stages are available in PTM. The first stage is how the document is to extract useful phrases from a text document. The second stage is then how to use these discovered patterns; this can be used to improve the effectiveness of a KDS.

i. Sequential Pattern Mining (PTM)
ii. Pattern Pruning
iii. Using discovered Pattern

Pattern Taxonomy is behave like a tree structure that illustrates the relationship between patterns extracted from a text collection. The aim of this step is to get rid of the worthless patterns using the closed relation. The main focus of this algorithm is deploying process, which consist of pattern discovery and term support evaluation. Assume that all text documents are split into paragraphs. So a given document d yields a set of paragraphs. Table.1. describes the set of paragraph and its terms. And also each paragraph is considered to coastline document, and this leads to the terms which can be intend from set of positive documents. Describe d = document; M= set of paragraphs; s=keyword. A sequential pattern [1] is used to get from collection of text travel permit and produce this pattern taxonomy model to represent relation between patterns being observed in the documents. Consider a small example for Pattern Taxonomy; Given Sequential Pattern p=(s1,s2,s3) in database. Lets contain minimum support main_sup=75%. In addition specified is the paragraph Ids where this pattern occurs. The pattern Taxonomy model method can apply to both the Positive and Negative documents. By which the user get required knowledge from guidance position of client profiles.

Table: 1 Set of Paragraph

Paragraph	Terms
dm1	s1,s2
dm2	s3,s4,s6
dm3	s3,s4,s5,s6
dm4	s3,s4,s5,s6
dm5	s1,s2,s6,s7

Eliminate bring to a close word and out of work words as well as subsequent words which occur many times and have less meaning. There are two phrases in this application phrase i.e. Training and Testing. Training phrase used to discover every repeated pattern from complete documents and prunes find the meaningless words and then calculate weights of the entire patterns. A centroid point is capable of patterns to be used to hold the demonstration of patterns from training set can be excerpt. The value of Pattern P is computed by following weight function:

$W(P)= (|\{da|\ da \in Dp, P \text{ in } da\ \}|)/(|\{db\ |\ db \in Dn, P \text{ in } db|)$

da and db denotes documents,
D denotes training set
Dp express positive documents
Dn expresses negative documents
Whereas Dp and Dn are the subsets of D.

Testing phrase is to find differences in positive/negative documents by the centroid obtained in training phrase by position each of them. The simple way to estimate similarity between documents and centroid from patterns. And also calculates summing weights of patterns which are available in the documents.

Table 2: Frequent Pattern Covering Set

Frequent Pattern	Covering Sets
{s3,s4,s6}	{dm2,dm3,dm4}
{s3,s4}	{dm2,dm3,dm4}
{s3,s6}	{dm2,dm3,dm4}
{s4,s6}	{dm2,dm3,dm4}
{s3}	{dm2,dm3,dm4}
{s4}	{dm2,dm3,dm4}
{s1,s2}	{dm1,dm5,dm6}
{s1}	{dm1,dm5,dm6}
{s2}	{dm1,dm5,dm6}
{s6}	{dm2,dm3,dm4,dm5,dm6}

In the similar method weights of patterns is designed you can also discover weights of individual terms in the document by a uncomplicated formula

$W(\ t)=\log \llbracket (d/(D-d))/((n-d)/((N-n)-\ (D-d\ ))) \rrbracket$

N is the total number of documents in training set,
D is the number of appropriate documents,
n is the amount of documents which contains t,
d is the quantity of relevant documents which contains t.

PTM uses closed sequential patterns, and display a certain extent improvement on the effectiveness. Two challenging issues for using patterns in Text Mining.

- The low support problem, for this given a topic, huge patterns are more specific for the topic, but which low support or low frequency. Decrease the minimum support, there are a lot of noisy patterns would be discovered.
- The misinterpretation problem that means the measures used in pattern mining cause misunderstanding of the document.

III. CLOSED SEQUENTIAL PATTERNS

Sequential patterns are the sequences which are defined by the user. The support exceeds the minimal support. Association rule mining use Apriori property was the prior algorithm can be used for sequential pattern mining. The transaction happened oftenly but it is not considered in a dataset. An association rule mines the intra transaction pattern and sequential patterns are to mine together into inner transaction pattern. ApproxMAP (APPROXimate Multiple Alignment Pattern Mining) [19] are grown to pride almost sequential patterns aggregate by many sequence, it covers many short patterns. An Apriori possession states that any sub patterns of frequent patterns should be frequent. Based on this, series of Apriori- like algorithm is proposed. GSP is an extension of Apriori model, uses “Generating-Pruning” method. PSP (Prefix Span) is another important method. It is fully based on “Generating-Pruning” principle. Applicants and frequent sequences are managed in more efficient structure in PSP than GSP. Another algorithm used for extracting sequential patterns is SPADE [9]. The main idea behind this method is a clustering of the frequent sequences based on their classification [24] of the candidate sequences and common prefixes.
Sequential pattern combined with no other sequential pattern that has the same support exactly is closed sequential pattern. Christo Ananth et al. [3] proposed a system, in which a predicate is defined for measuring the evidence for a boundary between two regions using Geodesic Graph-based representation of the image. The algorithm is applied to image segmentation using two different kinds of local neighborhoods in constructing the graph. Liver and hepatic tumor segmentation can be automatically processed by the Geodesic graph-cut based method. This system has concentrated on finding a fast and interactive segmentation method for liver and tumor segmentation. In the preprocessing stage, the CT image process is carried over with mean shift filter and statistical thresholding method for reducing processing area with improving detections rate. Second stage is liver segmentation; the liver region has been segmented using the algorithm of the proposed method. The next stage tumor segmentation also followed the same steps. Finally the liver and tumor regions are separately segmented from the computer tomography image. This method avert the spreading a sequence by expose the extension that is already included in sequence.

Closed sequential pattern is a frequent sequential pattern such that it is not included in another sequential pattern having exactly matching carry. A sequential pattern s=<t1; . . . ; tr> (ti elements of T) is an ordered list of terms. A sequence s1= <x1; . . . ; xi> is a subsequence of another sequence. s2<=y1; . . . ; yj>, is called s1 is sub-set of s2, if j1; . . . ; jy such that 1<= j1 < j2.. < jy <=j and x1=yj1; x2=yj2; . . . ; xi=yjy. Given s1 is sub-set of s2; they generally declare s1 is a sub pattern of s2, and s2 is a super pattern of s1.

IV. INNER PATTERN EVOLUTION

In a discovery model system, Effective Pattern Evolution is a methodology that is needed after pattern discovery phrase. The easiest way is to treat pattern atoms in a feature space to represent of the concept of a set of document, used in discovering patterns. The supports of terms d-patterns based on negative documents in the training sets. Because of low frequency problem, this limited the side effects of noisy patterns. It changes a pattern’s term support within the pattern; and this technique is called inner pattern evolution. Documents can be converted into diminish relevant or irrelevant categories based on a threshold. Pattern taxonomy mining [20], is used algorithm to improve the efficiency. SP Mining was proposed to find all closed sequential patterns, which uses Apriori property to reduce the searching space in order. In this section, discuss how to rationalize supports of terms within normal forms of d-patterns. The technique will be useful to reduce the side effects of noisy patterns because of the low-frequency problem. This technique is called inner pattern evolution here, because it only changes a pattern’s term supports within the pattern of particular documents. A threshold is usually used to classify documents into relevant or irrelevant categories. Using the d-patterns, the entrance can be defined naturally as follows:
Threshold (dp) = minp E DP (Σtw E β p support (t))

The shuffling of supports of terms d-patterns based on negative documents in the training set. Because of the low-frequency problem it reduces the side effects of noisy patterns. It changes only a pattern’s term supports within the pattern, this technique is called inner pattern evolution. There are two sorts of offenders: 1) a complete conflict offender which is a subset of A; and 2) a partial conflict offender which holds part of terms of A. This method is called inner pattern evolution because changes a pattern’s term supports within the pattern only. A threshold is used to classify documents into relevant or irrelevant categories. In order to reduce the noise, d-patterns are tracked and find out which pattern give rise to such an error. These patterns are offenders. There are two types of offenders complete conflict offend and Partial conflict offender the idea of updating patterns is explained as follows: Complete conflict offenders are removed from the discovered d-pattern at first. For partial conflict offenders reshuffling of the term support is carried out in order to reduce the effects of noise documents. Fig.2. shows the IPE methodology. This algorithm gives the better result and efficient updating of discovered pattern which is extracted from the text document.

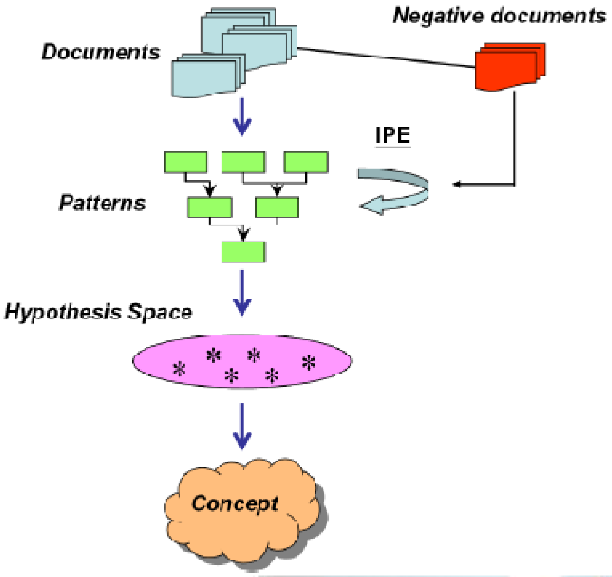

Fig.2. IPE Evaluation

The technique will be useful to reduce the side effects of noisy patterns because of low frequency problem. It only changes a pattern's term supports within the pattern. One possible solution to improve the efficiency of pattern based model is to reduce the dimensionality of the knowledge base.

V. PATTERN DEPLOYING METHOD (PDM)

A pattern deploying method for interpreting discovered pattern that provide a new method for weighting terms. The higher level patterns are closed sequential patterns that frequently appeared in paragraph. An efficient and effective way for using patterns in solving problems, especially for using large patterns. A pattern there is needed to identify the terms in the document and depict the pattern. All the terms are forming in the document. To overcome the low-frequency problem of specific long patterns for pattern deploying methods. The goal of through the use of a pattern composition operator, the goal of reserving the significant information embedded in specific patterns can be achieved the deploying patterns. When we deploy patterns to a common hypothesis space taken the pattern's support contained in the pattern discovery phrase. The big obstacle of pattern mining based approaches for text mining is has to effectively use both relevant and irrelevant feedback. In a pattern deploying method was proposed to adapted positive patterns. The improved effectiveness was not significant. The problem of redundancy and noises, another challenging issue for pattern based methods. And also deal with low frequency patterns.

A short pattern means normally with large support or called a highly frequent pattern is usually a general pattern. A large pattern is a low frequent pattern with small support could be a specific one. The Robust and effective pattern deploying techniques need to be implemented. The low capability of dealing with the mined patterns. There are several ways to utilize discovered patterns by using frequency. A pattern mining method that treated each found sequential pattern as a whole item without breaking them into a set of individual terms. For each positive document d€D^+, a set of patterns are discovered in order to be merged into a dedicated vector.

$$D_k = < (tk_1, nk1), (tk_2, nk2)\ldots\ldots (tk_m,nk_m)>$$

(t_{ki},n_{ki}) denotes an individual term

N_{ki}= |coverset ({t_{ki}})| total supports obtained from all patterns in dk.

Table.3. Example of set of positive documents consisting of Pattern Taxonomies

Document	Pattern Taxonomies	Sequential Patterns

D_1	$PT_{(1,1)}$
	$PT_{(1,2)}$
		{$(carbon)_4$,$(carbon,emiss)_3$}
		{$(air, pollut)_2$}

D_2	$PT_{(2,1)}$
	$PT_{(2,2)}$
		{$(greenhouse, global)_3$}
		{$(emiss, global)_2$}

D_3	$PT_{(3,1)}$
	$PT_{(3,2)}$
		{$(greenhous)_2$}
		{$(global, emiss)_2$}

D_4	$PT_{(4,1)}$
	$PT_{(4,2)}$
		{$(carbon)_3$}
		{$(air)_3$,$(air,antarct)_2$}

D_5	$PT_{(5,1)}$	{$(emiss,global,pollut)_2$}

For example, documents in Table3 can be represented by following vector:

D1 = <(carbon,2), (emiss, 1), (air,1), (pollut,1)>
D2 = <(greenhouse,1),(global,2),(emiss,1)>
D3=<(greenhou,1),(global,1),(emiss,1)>
D4=<(carbon,1),(air,2),(antarct,1)>
D5 =<(emiss,1),(global,1),(pollut,1)>

After completion of this patterns to merge the patterns using composition of operation. The composition of two patterns pattern 1 and pattern 2 can be processed. P_1□ P_2 = {(t,f_1+f_2) | (t,f_1) €p_1, (t,f_2) € p_2} U {(t,f) | t € (termset (p_1) U termset (p_2)) – (termset ()p_1) ∩ termset(p_2)), (t,f) € p_1Up_2}.The result of merging the documents with the deployed pattern that is the set of term weight pairs for each document is obtained.

D={(carbon,13/20),(emiss,67/60),(air,7/10),(pollut,8/15),
(greenhouse,7/12),(global,7/6),(antarct,1/4)}

Terms emiss and global are more likely to gain higher scores than the others. It gives the high appearance among sequential patterns.

VI. PATTERN DEPLOYING BASED ON SUPPORT (PDS)

The phrase of pattern discovers is taken into account when we deploy patterns. A probability function is also introduced to estimate the feature's significance. By using sequential pattern mining is pattern discovery technique can acquire a set of frequent sequential patterns SP for all documents d€ D^+. The following equation can support absolute discovered pattern,

$$\text{Support} :: SP \dashrightarrow [0,1]$$

$$\text{Support}(pi) = \frac{SP(pi)}{\sum pj \in SP\ (SP(pj))}$$

The composition operation in PDM Method is still available for the expanded forms of patterns in such a kind of format. To deploy a pattern there is need to identify the terms in the document and depict the pattern where all the terms are forming in the document. Pattern deploying methods are used to overcome the low-frequency problem of specific long patterns. Deploying patterns through the use of a pattern composition operator, the goal of reserving the significant information embedded in specific patterns can be achieved. The pattern‘s support obtained in the pattern discovery phrase is taken into account when we deploy patterns to a common hypothesis space. By using SP Mining algorithm can acquire a set of frequent sequential patterns [8] SP for all document d Є D+. Such that SP={p1,p2,p3,…pn}. The absolute support support (pi) for all pi Є SP is obtained as well. Normalize the absolute support of each discovered pattern based on the equation

Support (pi)= support (pi)/∑pjЄ SP support(pj)

Table.4. Sequential pattern Support values

Document	Sequential Patterns	Supports
D1	<carbon> <carbon, emiss> <air, pollut>	4/9 1/3 2/9
D2	<greenhouse, global> <emiss, global>	3/5 2/5
D3	<greenhouse>,<global, emiss>	½
D4	<carbon>,<air> <air, contract>	3/8 ¼
D5	<emiss,global,pollut>	1

Based on the above normalization Table.4, the expanded form of pattern pi can be represented using the following format:

Pi = <(ti,1,fi,1),(t1,2,fi,2),….,(ti,m,fi,m)>

Where *fi,j= support(pi)/m*

D-pattern mining algorithm is used to discover the D patterns from the set of documents. The efficiency of the pattern taxonomy mining is improved by proposing an SP mining algorithm to find all the closed sequential patterns, which is used as the well-known appropriate property in order to reduce the searching space. For every positive document, the SP Mining algorithm is first called giving rise to a set of closed sequential patterns. The main focus is the deploying process, which consists of the d-pattern discovery and term support evaluation. Christo Ananth et al. [6] proposed a system, this system has concentrated on finding a fast and interactive segmentation method for liver and tumor segmentation. In the pre-processing stage, Mean shift filter is applied to CT image process and statistical thresholding method is applied for reducing processing area with improving detections rate. In the Second stage, the liver region has been segmented using the algorithm of the proposed method. Next, the tumor region has been segmented using Geodesic Graph cut method. Results show that the proposed method is less prone to shortcutting than typical graph cut methods while being less sensitive to seed placement and better at edge localization than geodesic methods. This leads to increased segmentation accuracy and reduced effort on the part of the user. Finally Segmented Liver and Tumor Regions were shown from the abdominal Computed Tomographic image. In term-based approaches, the evaluation of term weights (supports) is based on the distribution 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. Various mechanisms in each step are proposed and evaluated for fulfilling the PTM model. A pattern deploying method based on support (PDS) is developed to deal with the discovered patterns in proper ways and provide suitable solutions for using these patterns.

VII. EXPERIMENTAL RESULTS

The most popular used dataset currently is RCV1, which includes 806,791 news articles for the period between 20 August 1996. These documents were formatted by using a structured XML schema. Using TREC filtering method provided two groups of topics. Totally 50 topics each like that training set and testing set. Documents in both sets are negative and positive. Positive documents means relevant to the assigned topic, otherwise negative. Fig.3. shows all experimental models use title and text of XML documents only. The title is shown as a paragraph as the one in text which consists of paragraphs. Most of the standard measures based on precision and recall are used. The precision is the fraction of retrieved documents that are relevant to the topic. The recall is the fraction of relevant documents that have been retrieved. The first precision of K returned top-K. K value set as a 10. The breakeven point is used to provide another measurement for performance evaluation. The breakeven measure has been frequently used in common information retrieval evaluations.

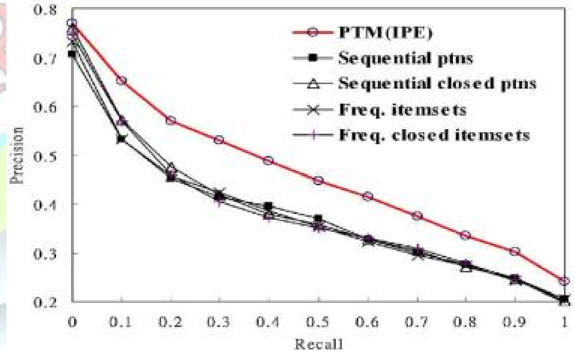

Fig.3. Comparing PTM with Data mining Methods for first 50 Topics.

$$F_{\beta\text{-measure}} = \frac{(\beta^2+1)*precision*recall}{\beta^2 *precision+recall}$$

Where β is a parameter giving weights of precision and recall and can be the relative degree. The value of $F_{\beta\text{-}1}$ is equivalent to the breakeven point when precision and recall. The $F_{\beta\text{-}1}$ measure is greater or equal to the value of breakeven point.

Table.5. Comparison for Pattern Mining Methods

Method	Top-20	Breakeven point	MAP	$F_{\beta-1}$	IAP
PTM	**0.493**	**0.429**	**0.441**	**0.440**	**0.466**
Rocchio	0.416	0.392	0.391	0.408	0.418
Sequential closed pattern	0.406	0.353	0.364	0.390	0.392
SVM	0.447	0.409	0.408	0.421	0.434
TFIDF	0.321	0.321	0.322	0.355	0.348
Frequent Item sets	0.412	0.352	0.361	0.386	0.384
Frequent closed itemsets	0.428	0.346	0.361	0.385	0.387
CBM	0.448	0.409	0.415	0.423	0.440
CBM pattern Matching	0.329	0.282	0.283	0.320	0.311

The breakeven point and $F_{\beta\text{-measure}}$ both are single valued reason and reflect the performance over the Fig.4 for the whole documents. While we using these measures need to use more figures to evaluate the system as a whole. Interpolated Average Precision (IAP) is introduced and has been adopted in several research works. Mean Average Precision (MAP) is used in our evaluation which is calculated by measuring precision at each relevance document first, and averaging precisions over all the topics.

VIII. COMPARISON OF PTM (IPE) VERSUS OTHER PATTERN MODELS

The number of patterns used for training by each method. The total number of patterns is estimated by accumulating the number of each topic. PTM is the method that utilizes the least amount of patterns for concept learning compared to others. The efficient scheme of pattern pruning is applied to the PTM method. The classic methods such as Rocchio, prob and TFIDF adopt terms as patterns in the feature space; they use much more patterns than the proposed PTM method and slightly less than the sequential pattern mining method.

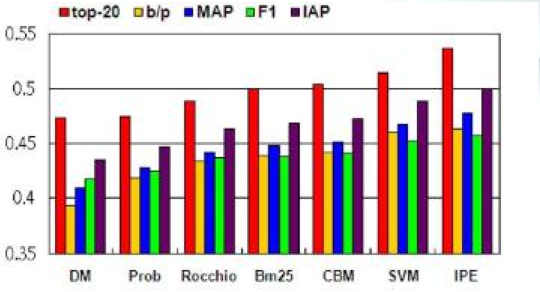

Fig.4. Comparison of PTM (IPE) and other major models in five measures for 100 topics

The results supports the superiority of the PTM method and highlights the importance of the adoption of proper pattern deploying and pattern evolving methods to a pattern based knowledge discovery system. All four pattern mining methods achieve similar results. The PTM is much smoother than those for pattern mining methods as there is no complex fluctuation on it. Pattern mining based methods can improve the performance in the low-recall situation. The PTM achieves an outstanding performance for text mining by comparing with the up-to-date data mining based methods, the concept based methods and BM25 and SVM models.

IX. CONCLUSION

Many text pattern mining techniques have been proposed for fulfilling various knowledge discovery tasks. However, using these discovered knowledge (or patterns) in the field of text mining is complicated and useless. The reason is that some useful long patterns with high specificity inadequacy in support (i.e., the low-frequency problem). Discussions based on that not all frequent short patterns are useful. Hence, the patterns are derived misconception from the techniques of text mining lead to the ineffective performance. This research work, a pattern discovery technique has been proposed to overcome the low-frequency and misinterpretation problems for text mining. One possible solution to improve the efficiency of pattern taxonomy-based model is to reduce the knowledge base extension of the feature space in the knowledge base. The result of this research is to explore the strengths and weaknesses of a discover new patterns and pattern based approach for text classification. Based on the experimental results, there are some pattern mining methods are to be compared and find Pattern Taxonomy method is efficient than other methods.

References

[1]Hye-Chung Kum, Joong Hyuk Chang, and Wei Wang: “Sequential Pattern Mining in Multi-Databases via Multiple Alignments. IEEE Trans on Data Mining Knowledge and Discovery. 12(2-3): Pp: 151-180, 2006.

[2]F.Sebastiani. “Machine learning in automated text categorization.” ACM Computing 34(1): Pp: 1-47, 2002.

[3] Christo Ananth, D.L.Roshni Bai, K.Renuka, A.Vidhya, C.Savithra, “Liver and Hepatic Tumor Segmentation in 3D CT Images”, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Volume 3,Issue-2, February 2014,pp 496-503

[4]R. Agrawal and R. Srikant. “Fast algorithms mining association rules.” In Proc.of the VLDB Conference, Santiago, Chile, September 1994.Expanded version available as IBM Research Report RJ9839, Pp:267-277,June 1994.

[5]Han, J. and Kamber, M. “Data Mining Concepts and Techniques.” 3rd edition, University of Illinois at Urbana-Champaign, Morgan Kanufmann publishers,Pp: 348-408, 2006.

[6] Christo Ananth, D.L.Roshni Bai , K.Renuka, C.Savithra, A.Vidhya, “Interactive Automatic Hepatic Tumor CT Image Segmentation”, International Journal of Emerging Research in Management &Technology (IJERMT), Volume-3, Issue-1, January 2014,pp 16-20

[7]J. Han and J. Pei. “Mining frequent patterns by pattern-growth: methodology and implications.” SIGKDD Explore. Newsl. Pp:14-20, 2000.

[8]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:11-12, 2000.

[9]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.

[10]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.

[11]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.

[12]R. Sharma and S. Raman. Phrase-based text Representation for managing the web document. In Proceedings of the International Conference on Information Technology: Computers and Communications (ITCC), Pp: 165–169, 2003.

[13]S-T. Wu, Y. Li, Y. Xu, B. Pham, and P. Chen.Automatic pattern-taxonomy extraction for Web mining. In Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligent (W104), Pp: 242–248, 2004.

[14]J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. In Proceedings of ACM-SIGMOD 2000, Pp: 1–12, 2000.

[15]Kavitha Murugeshan, Neeraj RK “Discovering Patterns to Produce Effective Output through Text Mining Using Naïve Bayesian Algorithm” IJITEE ISSN: 2278-3075, Volume-2, Issue-6, Pp: 28-29, May 2013.

[16]Y.Li and N.Zhong. “Mining ontology for automatically acquiring web user information needs.” IEEE Transaction on Knowledge and Data Engineering, 18(4): Pp: 554-568, 2006.

[17]S. Shehata, F. Karray, and M. Kamel. “A concept-based model for enhancing text categorization”. In Proc. Of KDD’07, Pp: 629–637, 2007.

[18]“Automatic Pattern-Taxonomy Extraction for Web Mining” Sheng-Tang Wu Yuefeng Li Yue Xu Binh Pham Phoebe Chen* IEEE Conference, Pp: 67-90, 2010.

[19]R. Agrawal, and R.Srikant, “Mining sequential patterns,” Proceedings of Int. Conf. on Data engineering (ICDE’95), Taipei, Taiwan, Pp: 3-14, 1995.

[20]Charjan.D, Mukesh.A “Pattern Discovery For Text Mining Using Pattern Taxonomy”, IJETT, Pp: 128-190, 2013.

[21]Radhakrishnan.A, Kurian.M “Efficient Updating of Discovered Patterns for Text Mining: A Survey” IJCSNS Pp: 329-443, 2013.

[22]Sahaphong.S and Boonjing.V” IIS-Mine: A new efficient method for mining frequent itemsets”, MIJST, Pp: 267-310, 2012.

[23]Warnars.S “Mining Frequent Pattern with Attribute Oriented Induction High Level Emerging Pattern (AOI-HEP)”, IEEE(ICoICT), Pp:87-95, 2014.

[24]Urmila.M “Pattern-Based Text Mining Method For Classification of Research Proposals”, IJRCCT, Pp: 24-67, 2014.

[25] B. Liu, Y. Dai, X. Li, W. S. Lee, and P. S. Yu, “Building text classifiers using positive and unlabeled examples,” ICDM03, 2003, Pp: 179- 186,1995.

ABOUT AUTHORS

S.Brindha received B.Sc degree in Physics 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 3 years of teaching experience and 5 years of Technical Experience in Hash Prompt Softwares Pvt. Ltd.
Currently She is doing her Ph.D computer Science in Erode Arts and Science College. Her Research area includes Data Mining and Text Mining.

K.Prabha received B.Sc Computer Science and M.Sc Computer Science Degree from Bharathiar University, Coimbatore and M.Phil in Periyar University, Salem. She received thePh.D degree in Computer Science at Bharathiar University. She has 7 years of teaching experience. She was working as Assistant Professor of Computer Science in Erode Arts and Science College,
Erode, Tamilnadu, India. She is working as a Associate professor in PG Extension Center, Periyar University, Dharmapuri, Tamilnadu, Erode. Currently she is Guiding 1 M.Phil Scholar and 1 Ph.D Scholar. She published around 15 research papers in national and international journals and conferences. Her research interests include Network Security and Data 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 25 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 50 M.Phil research Scholars in various fields and guided two Ph.D Scholars. Currently he is Guiding 5 M.Phil Scholars and 6Ph.D Scholars. He is a member of Board studies of various Autonomous Colleges and Universities. He published around 40 research papers in national and international journals and conferences. His current research interests include Image processing and Network Security.