A Survey Paper For Finding Frequent Pattern In Text Mining

Ms. Sonam Tripathi[1] , Asst Prof. Tripti Sharma[2]

***Abstract*— In text document, huge data mining techniques have been used for mining useful pattern. Text mining can be used to extract the large database or datasets from the document or paragraph. The text mining is used on existing term-based approach and produces the problem of polysemy and synonymy. The frequent pattern based approach performs better than the term-based ones, but sometime this experiment is not working. This paper includes the process of pattern deploying and pattern evolving and improves the effectiveness of discovering patterns for finding useful information**

***Index Terms*— Data mining; text mining; frequent pattern mining; term-based method; pattern evolution.**

I. INTRODUCTION

Text mining is a technique that is used to find useful information from large amount of data sets. Data mining has rules called as frequent pattern and association rule that is important for finding frequent patterns. The apriori based algorithm and tree structure-based algorithms are used in frequent pattern mining [13]. We are using a tree structure-based algorithm, this algorithm follows a test approach an test support frequencies only. Example are FP tree and FR growth tree [13].

In the last decade, data mining has been proposed different knowledge tasks. These tasks include sequential pattern mining, maximum pattern mining, association rule and closed pattern mining. The synonymy and polysemy method are creating a problem in term-based method [5], [8], [12]. A word that share the same meaning in other word that is called synonymy and a word that has 2 or more meanings that is called polysemy [5].

This paper proposes a temporal text mining approach for frequent pattern mining . Temporal text mining combines data mining techniques and extracting information upon texting repository [13]. The sequences of events from the sets of documents are extracted in order to track the past events effectively. Associated with the given document set are constructed the optimal decomposition of the time period. The notion of the compressed level decomposition is introduced where each sub-interval consists of consecutive time points having identical information content [3]. Several documents are defined based on the information computed as document sets are combined.

***Manuscript received March, 2014*.**

*Ms. **Sonam Tripathi**, Computer Science & Engineering, RCET Bhilai Bhilai, India.*

***Asst Prof Tripti Sharma,**,Computer Science & Engineering,RCET Bhilai, India.*

Text mining is the discovery of important knowledge in text mining. It is a challenging issue to find Knowledge to help users to find what one wants. Information Retrieval (IR) provided many term-based methods to solve this problem [1], [3], [12].

There are two main issues of pattern-based method: low frequency and misinterpretation (miss understanding) [12]. A highly frequent pattern is usually a specific pattern of low frequency. Many noisy patterns are discovered, if we decrease the minimum support. Misinterpretation means the measures used in pattern mining (e.g., "support" and "confidence") turn out to be not suitable in using discover patterns to answer what users want. In text document, the difficult problem is how to use discovered patterns to accurately evaluate the weights of useful knowledge [3], [12].

The FP growth algorithm has three advantages: First, it scans database two times and decreases computational cost. Second, generation of candidate item-sets is removed in the FP tree algorithm. Third, it reduces the search space by using the divide and conquers method. The FP growth algorithm has a disadvantage. It can not used in incremental mining because when new transactions are added to the database [5], FP tree requires updating in the data sets and whole process are repeated after this. This problem has received attention from researchers in data mining and information retrieval(IR) communities.

II. LITERATURE REVIEW

Earlier papers were describing the method for discovering frequent pattern by using pattern taxonomy models and various techniques as illustrated and discussed here .

Many types of text representations have been proposed in the past. A well known one is the bag of words that uses keywords (terms) as elements in the vector of the feature space. the weighting scheme is used for text representation in Rocchio classifiers has proposed by X. Li and B. Liu, in Learning to Classify Texts Using Positive and Unlabeled Data, in the year 2003 [1], [7], [12], [14]. Various weighting schemes for the bag of words representation approach were given by K. Aas and L. Eikvil [12], [14].

The combination of unigram and biagrams was chosen for document indexing in text categorization and evaluated on a variety of feature evaluation functions (FEF) has proposed by M.F. Caropreso, S. Matwin, and F. Sebastiani in the year 2000.

Data mining techniques have been used for text analysis by extracting occurring terms as descriptive phrases from document collections [14]. The effectiveness of a text mining a system using phrases as a text representation showed no significant improvement. The likely reason was that a phrase-based method had "lower consistency of assignment

and lower document frequency for the terms" as mentioned in the article An Evaluation of Phrasal and Clustered Representation on a Text Categorization Task in the year 1992 [5], [12], [21].

Discoves frequent pattern mining algorithm include Agrawal.R & Batra.M (2013), Bhaskar.R & Srivatsan. L et al(2010),Pipanmaekaporn.L(2013),Radhakrishnan.A& Kurian.M(2013) at mining sequences and trees respectively.

a. Knowledge Discovery

Knowledge discovery is the process of nontrivial extraction of information from large database information that is implicitly present in data mining, previously unknown useful for users [14], [15].
The field of knowledge discovery has developed since the year of 1980s. The research trend in knowledge discovery belong the following issues:-
a.1. Mining association rules efficiently.
a.2. Mining object oriented databases.
a.3. Mining multimedia data.
a.4. Mining distributed and heterogeneous databases.
a.5. Text mining.
a.6. Knowledge discovery in semi-structure data.

b. Association Analysis

Association rule is discovered interesting patterns from a given dataset. They are processed with data mining. The association rule mining is used in the market basket analysis, which searches for relationships between shoppers and items bought. Two common measures of rule usefulness are rule support and confidence. It is calculated by the percentage of task relevant data transaction for which a pattern is recognized as true [4], [13].The association rule approaches has two phase: support phase and confidence phase. The support and confidence of the rule A=>B can be expressed as the following equation.
b.1. Support
The rules with support sup in T, the transaction data set if the sup% of transactions is A ∪ B.
b.2. Confidence
The rule with confidence in T if the conf% of transactions that is A also contain B.

$$\text{support}(A \Rightarrow B) = P(A \cup B)$$
$$\text{confidence}(A \Rightarrow B) = P(B \mid A)$$

c. Text Mining

Most work in knowledge discovery and data mining was concerned with structured databases. The text article has a large portion of the available data appears in the collection. Text mining is used for all text to extract useful information by finding potential pattern from large quantities of text. It combines many terms such as information retrieval, information extraction, machine learning, text categorization and data mining [1], [11]. Considering the sequence of the words in a transaction is vital for finding language patterns, in text mining.
c.1. Keyword based representation.
c.2. Phrase based representation.

III. PROPOSED METHODOLOGY

In this work we will try to solve the problem of pattern discovery using a term-based method with a numerical representation of the frequent pattern tree. As the method of pattern discovery introduces the problems of polysemy and synonymy This paper produces a solution to overcome from above problem and produces better and efficient solutions.

- A KDD process based on Pattern taxonomy model is to be proposed.
- In the Pattern taxonomy model used in sequential pattern and closed sequential pattern in the state-of-art data mining techniques.
- A scalable pattern taxonomy model is develop with the capacity of concept adjustment.
- Pattern Deployment strategies are provide to increase the effectiveness of the PTM.

a. FP TREE

In which all items are arranged in descending order of frequency and thresholds i.e. a tree structure. The frequent items can be mined using FP-growth [8], [13], [14].

a.1. Creation on FP-Tree

A transactional database is divided in 2 categories that is transaction id and the list of items and then scan the entire database. Collect the threshold of the items in presenting database, then sort the items in decreasing order on their no. of occurrences [8].

Now, the transactional databases are scanned. Scanning will be started from root node. Further add the transaction database one after another as prefix sub-trees the root node. This process is repeated until all the transaction have been include the frequent-pattern tree. A header table is constructed that consist of items and counts.
Consider the transactional database in below table with 5 transactions.
Table 1. Example of Transactional Databases

Tran.ID	Items
T1	A,B,F
T2	A,C,D,E
T3	E,F,H,I,B
T4	A,B
T5	C,E

The frequent item sets for the above database is given in Table 2.

Table 2. Frequent Item sets for the transactional database in Table 1

Items	Count
A	3
B	3
C	3
D	1
E	3
F	2
H	1
I	1

The items that do not meet the minimum threshold has been eliminated. The frequent item sets that support the minimum support threshold is given in Table 3 [12].

Table 3. Frequent Item list for the Transactional Database that Support Minimum Threshold

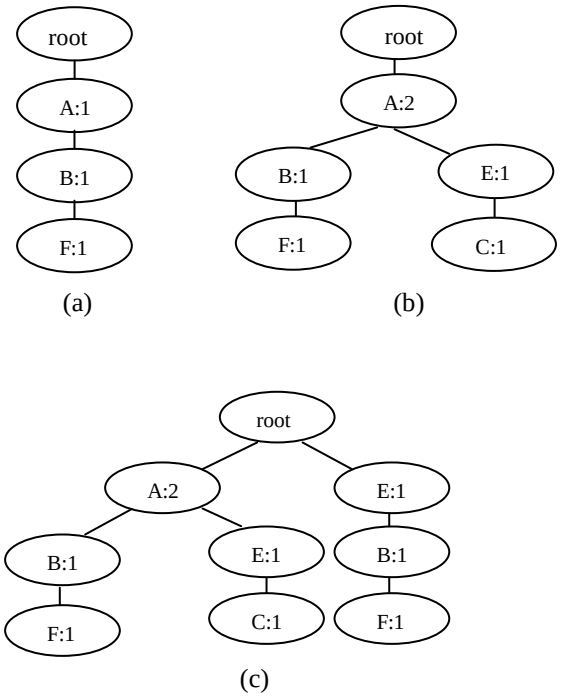
Item	Count
A	3
B	3
E	3
C	2
F	2

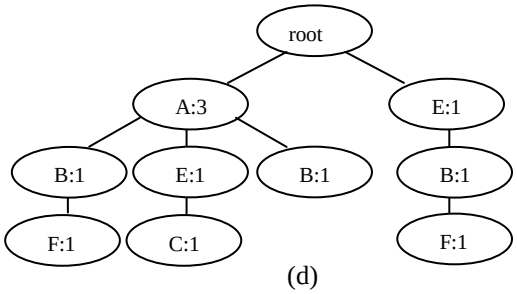
The transaction database according to the frequent item list is given in Table 4.

Table 4. Sorted and Eliminated Transaction of the Database in Table 1.

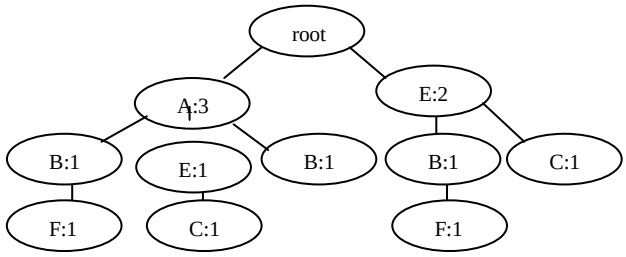
Tran.ID	Items
T1	A,B,F
T2	A,E,C
T3	E,B,F
T4	A,B
T5	E,C

Diagram





(d)



(e)

Figure 1:- Steps in creating the FP-Tree

a.2. Finding Frequent Pattern on FP-Tree

The frequent pattern can be mined using an iterative approach FP-growth, after construction of the FP-tree. This approach is shown in the header table in the above figure . The FP growth selects the items that support the minimum support [4], [12], [22]. It removes the infrequent items from the prefix path of an existing node.

a.3. Advantages and Disadvantages

It does not generate any candidate key, that is the advantage of frequent pattern [8], [12]. It suffers from the issue of special and temporal locality issues, that is the disadvantages of frequent pattern [8].

b. PROTOTYPE OF PATTERN TAXONOMY MODEL

Knowledge discovery includes a lot of data mining method that have been proposed for conquering related challenges in different field, especially the domain of super basket data and telecommunication data. It is difficult to find a suitable example, that can implement these data mining techniques in the area of text mining, which is analysed by the use of information Retrieval-related method [9], [13].

b.1. Pattern Taxonomy Model

Two main stages are available in PTM. The first stage is how to extract useful phrases from a text document. The second stage is then how to use these discovered patterns to improve the effectiveness of a KDS [1], [12], [13].

i. Sequential Pattern Mining (SPM).
ii. Pattern Pruning.
iii. Using Discovered Pattern.

b.2. Finding Non-Sequential Pattern

The essential definition of NSPM is described as follows: Let T = {t1,t2,t3,…,tn} be a set of distinct terms. A non-sequential pattern p is a subset of p.

c. APRIORI ALGORITHM

Apriori algorithm was introduced for frequent item set and Association rule mining. Apriori algorithm is proposed by R. Agrawal and R Srikant in 1994. Apriori is an important development, in history of association rule mining. Apriori algorithm overcomes the limitations of AIS algorithm which generate too many candidate itemsets and most of them are infrequent [4],[13][23].

Apriori algorithm works on two concepts-:
c.1. Self Join- Two item sets having one and only one different item can be joined.
c.2. Pruning- Candidates containing non-frequent subsets are removed.

This process is repeated until frequent item set or candidate item set. The transaction database is scanned first time for candidate itemset which consist one item set and support count is calculated. After these 1-candidate itemset are pruned by eliminating those itemsets that has an item count less than the user-defined threshold (for example threshold=40%). In second phase database is scanned again for a generation in 2-candidate Itemsets which consist of two items, then again pruning is done according to Apriori [4], [5], [13] .

There are two drawbacks of this algorithm First is the generation of the complex candidate item set which requires large memory and execution time and the second problem it requires lots of database scans for candidate generation.

d. INNER PATTERN EVALUTION

In a discovery model system, Effective Pattern Evolution is a methodology that is needed after pattern discovery phase. 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 [5], [9], [12]. Inner pattern evolution and pattern deploying these techniques are includes an effective pattern discovery model.
In the training set, we discuss about the shuffling of the supports of terms d-patterns based on negative documents in the training sets. Because of low frequency problem, this reduces the side effects of noisy patterns. It changes a pattern's term support within the pattern, this technique is called inner pattern evolution. Documents into relevant or irrelevant categories based on a threshold [23].

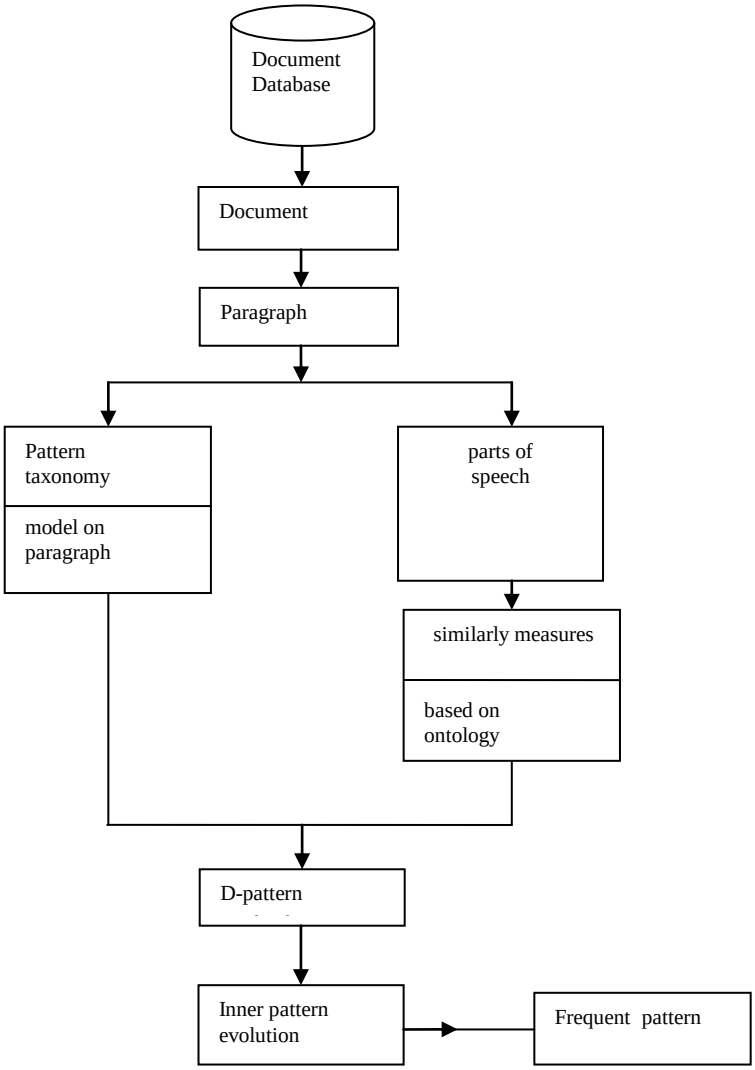


Figure 2-: shows To find the frequent pattern used in PTM, D-pattern and Inner pattern evolutio

IV. CONCLUSION

In the last decade, many data mining techniques have been proposed. Frequent item set mining, closed pattern mining, sequential pattern mining, association rule mining and closed pattern mining these all techniques are used in data mining frequent pattern techniques. The pattern deploying and pattern evolving these two techniques are used in proposed technique. In this research work solves the problem of low frequency and miss interpretation in pattern discovery techniques . Some research work uses phrases rather than individual work. The proposed model outperforms term-based state-of-the-art models, such as BM25 and SVM-based model not only concept-based model.

REFERENCES

[1] Agrawal.R, Batra.M "A Detailed Study on Text Mining Techniques" IJSCE 2013.
[2] Bhaskar.R, Srivatsan.L, Smith.A, Thakurta.A "Discovering frequent patterns in sensitive data" ACM, 2010.
[3] Gangarde.R , Kohle.V "Effective Pattern Discovery by Cleaning Patterns with Pattern Co-occurrence Matrix and PDCS Deploying Approach" IEEE , 2014.

[4] Goswami.D, Chaturvedi.A, Raghuvanshi.C "An Algorithm for Frequent Pattern Mining Based On Apriori" IJCSE 2010.
[5] Inje.B, Patil.U "Operational pattern revealing technique in textmining", IEEE Students' Conference on Electrical,Electronics and Computer Science, 2014.
[6] Jadhav.J, Ragha.L, Katkar.V "Incremental Frequent PatternMining", IJEAT2012.
[7] Joshi.S, Jadon.R, Jain "An Implementation of Frequent Pattern Mining Algorithm using Dynamic Function", IJCA 2010.
[8] Sasireka.K, Raja.K, Kiruthinga.G "A survey about various data structures for mining frequent patterns from large database", IJRRIS Science Academy Publisher, September 2011.
[9] Mythili.K, Yasodha.K "A Pattern Taxonomy Model with New Pattern Discovery Model for Text Mining" IJSAIT 2012.
[10] Mishra M, Choubey M "Discovery of Frequent Patterns from Web Log Data by using FP-Growth algorithm for Web Usage Mining" IJARCSSE 2012.
[11] Charjan.D, Mukesh.A "Pattern Discovery For Text Mining Using Pattern Taxonomy", IJETT 2013.
[12] Zhong.N, Yuefeng.L "Effective pattern discovery in text mining" IEEE TRANSACTION ON KNOWLEDGE AND DATA ENGINEERING, VOL,NO.1, January 2012.
[13] Mabroukeh.N and Ezeife.C "A Taxonomy of Sequential Pattern Mining Algorithms" ACM Computing Surveys, Vol. 43, No. 1, Article 3, Publication date, November 2010.
[14] Pipanmaekaporn.L "Feature Discovery in Relevance Feedback Using Pattern Mining", IEEE 2013.
[15] Radhakrishnan.A, Kurian.M "Effective Pattern Matching Approach for Knowledge Discovery Application" IJARECE Volume 2, Issue 2, February 2013.
[16] Radhakrishnan.A, Kurian.M "Efficient Updating of Discovered Patterns for Text Mining: A Survey" IJCSNS 2013.
[17] Kumar.R, Verma.R "Classification Algorithms for Data Mining:A Survey", IJIET 2012.
[18] Sahaphong.S and Boonjing.V" IIS-Mine: A new efficient method for mining frequent itemsets", MIJST 2012.
[19] Shyam Sudar Meena "Efficient Discovery of Frequent Patterns using KFP-Tree from Web Logs" IJCA 2012.
[20] Fakhrahmad.S, Dastghaibyfard.G "An Efficient Frequent Pattern Mining Method and its Parallelization in Transactional Databases", JISAE 2011.
[21] Urmila.M "Pattern-Based Text Mining Method For Classification of Research Proposals", IJRCCT 2014.
[22] Warnars.S "Mining Frequent Pattern with Attribute Oriented Induction High Level Emerging Pattern (AOI-HEP)", IEEE(ICoICT) 2014.
[23] Ghosh.S, Biswas.S, Sarkar.D, Sarkar.P "Mining Frequent Itemsets Using Genetic Algorithm" ,IJAIA 2010.

About Authors:

Ms.Sonam Tripathi received the B.E. degree from Chhattigarh Swami Vivekanand Technical University, Bhilai (C.G.) India in Computer Science & Engineering in the year 2012. She is currently pursuing M.Tech. Degree in Computer Science Engineering with specialization in Computer Science & Engineering from CSVTU Bhilai (C.G.), India. Her research area includes Data Mining and Text Mining etc.

Ms. Tripti Sharma is currently Assistant professor in Department of Computer science & Engineering RCET, Bhilai (C.G.) India. She completed her B.E and M.Tech. in Computer Science and Engineering Branch. Her research area includes Data mining, Image processing, Computer Network, AI & NN etc. She has published many Research Papers in various reputed National & International Journals, Conferences, and Seminars.

