Effective Pattern Discovery for Text Mining using AprioriDP and LDA Algorithm

S.Vijaya
Assistant Professor,
Department of Information Technology,
Kg College of Arts and Science ,
Coimbatore.
s.vijaya@kgcas.com

Abstract: Data mining is the process of analyzing data from the different large amount of data to increasing the cost, revenue, cuts and useful information from the databases.Most of the text mining process is adopted in the normal traditional language text.Text mining process can work with unstructured or semi-structured data to convert numerical values which can be better solution for structured data on data mining techniques. In Most existing mining methods cannot simply the sturtured data and problems in polysemy and synonymy. In this paper, we proposed the combination of AprioriDP and LDA Algorithm uses the context variable in pattern mining (feature extraction) and to eliminate the candidate generation is proposed. This paper mainly focused on developing the effective mining algorithms for discovering patterns from large volume of data. To improve the effectiveness of using and updating discovered patterns for finding relevant and filtering information. The experimental results improve the performance and to discovering the effective results in mining.

1. INTRODUCTION

Data mining is process of extracting useful data from the large volume of data in the databases. To increase the evaluation of data with useful information from extracting data.Many of the mining software is used in the collection of data. But Data mining is one of a number of analytical tools used for analyzing data from the databases. It allows users to analyze data from many different dimensions or angles, categorize it, and structured data with summarize the linkedrelationships are identified[1]. Technically, data mining is the process of finding correlations or patterns among more number of fields in large relational databases. In these techniques are used to analysis, summarization, and categorization of all set of items that support transaction in minimum level. This support is an itemsets are including the extraction, analysis, etc.In this support used in large set of itemsets and all others are called small itemsets[2].Large number of itemsets are classified with desired rules[3]. So we are using straight forward algorithm for this task. Most of these techniques due to lack of space, we do not discuss this one of problem further, but refer the effective efficient for a fast algorithm. Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD)[4], an interdisciplinary subfield of computer science, is the computational process of discovering patterns in large data sets to connect with the algorithms and methods of KDD, artificial intelligence,machinelearning,statistics, and database systems. So we have to discuss about that data mining process is to extract the useful information from a dataset and classify the understandable structure of each data set to mining the data easily.

But in the analyzing data is the first step of datamining process to management the dataset. Next to managing the data aspects, and preprocess the data, model of transactions, inference considerations, interestingness advantages, complexity considerations, post processing of discovered patterns, visualization, online updating ,selection process, database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, classification rules, visualization, and online updating.Feature extraction and selection, pattern evaluation.

Text mining is the important research process in the data mining.To discovering the knowledge of information in text documents.Its deals with extracting information from large database. But it is a challenging issue to find structured data with retrieving the information that user require relevant efficiency. Text mining different fields which on information retrieval (IR), machine lerarning using SVM and knowledge discovery text mining in databases. In the beginning, information Retrieval (IR) is searching and retrieving a set of documents from a document collection to response to a search query. In IR process provided many approches to solve this problem such as Rocchioand probabilistic models, rough set models, BM25 and support vector machine (SVM) based filtering modes[5],[6].

Several applications are used the IR process and search engines like google, Bing, Facebook, Twitter, Yahoo, LinkedIn.To extracting explicit and implicit concepts and relationship between structured data using natural language processing (NLP), statistical modeling and machine learning techniques can be challenging, however, because the natural language text is often inconsistent. It contains ambiguities caused by inconsistent syntax and semantics, including words slang, language specific to vertical industries and age groups, double entendres and sarcasm.

Pattern mining has been extensively in data mining concepts for many years. Pattern mining discovering patterns and reduces their influence of pattern and improve the evaluation specific in documents. To examine this technique closed patterns in text mining was very useful and potential for improving the performance of text. However, using the discovered knowledge (or patterns) in the field of text mining is difficult and ineffective. Different algorithms such as Apriori algorithm, spade, PrefixSpan, SLPMinerare used.Discovering knowledge patterns in text mining are difficult and ineffective. The reason is that long support itemsets with high specificity lack of support (i.e., low frequency problem). So that evaluation technique used to improvew the performance of term based approach in tecxt mining.

2. RELATED WORK

As the volume of information is available on the social sites users are want better tool need for growing the information from collection of datas. Text categorization is an important to managing many task and this process is very time consuming and cost oriented also[1]. Data mining techniques have been extract and analyzed with phrases from the document collections[4].

The documents are classified and representation by keywords as elements in the vector of the feature space.The ability to search for keywords or phrases in the database and marginally collects that report only user which word is mostly specified in that text preference looking for.Term based ontology methods useful ina knowledge domain create ontologies and enchance the process of search and retrieval on the document.

For example, hierarchicalrelationship clustering was used to determine synonymy and hyponymy relations between keywords. Also, thepattern evolution technique was introduced to improve the accuracy and effective results in mining.Pattern mining process is used in data mining for many years.A variety of efficientalgorithms such as Apriori, PrefixSpan, FP-tree, SPADE, SLPMiner, and GST have been proposed.[11].

Caropreso, S. Matwin et al[7], chosen to work by substitution because, unlike in IR, in pattern dimensionality of the feature space is an important, and because of this any comparison between different representation schemes is significant only if the numbers of features used are the same.Cortes et al[8], says the idea behind the support-vector network was previously implemented for the restricted case where the data can be widespread separated without faults occur. We here extend this result to non-separable training data.

In this model[9], several methods are discussed in the similarity ofterms and documents is determined by the overall pattern of word usage in the entire collection. So that documents can be similar to each other, information retrieval to help overcome the problem and improve the performance.A knowledge discovered in terms, based on the classification structure data the number of dimensions for the reduced space.

Based on [10], Fully automatic extraction of Web page structures andsemantic contents can bedifficult given the current limitations on automated natural-language parsing. This method evaluated objectively with informally.

These research workshave mainly focused on extracting and developing efficient mining algorithms for discovering patterns from a large volume of data collection.However, searching for useful information and interesting patterns and rules werestill an open problem. In the field oftext mining, pattern mining techniques can be used to findvarious text patterns, such as sequential patterns, frequentitem sets, co-occurring terms and multiple grams, for buildingup a representation with these new types of features. The proposed approach is how to effectively dealwith the large amount of discovered patterns.

The pattern deploying and the pattern evolving used in existing methods will not work for elimination of unwanted candidates and it will not be a context based methods. The problems of existing method are as follows, Classification of rules for discovering the terms would not benefit from changing the order of the sentiment-related terms when it exceeded the threshold value.Classification results provide same probability results for both old and new products. (For example: Nokia 1100 and Nokia Lumina). No methods are produced to work on Ironic Phase (that is Giving rumor data as actual data).No methods clearly identified the domain in all phases. (For example: Predicting Obama rules are good or bad, In some situation or for some people, rule is good, sometimes it is bad, But the system will show only Goodor Bad). Another one example is Diabetes diagnosis dataset contains attributes including class attribute. Numbers of instances are included and where last attributes shows sick or healthy. Sub-domain Analysis should be concentrate to improve the accuracy of sentiment analysis.Time-to-Time analysis should be calculated efficiently to improve the accuracy of sentiment analysis.More efficient sentiment calculation algorithm to enhance the accuracy of judging the sentiment from the reviews.

3. PROPOSED STUDY

The proposed system is the combination of **Apriori** and LDA Algorithm uses the context variable in pattern mining (feature extraction) and to eliminate the candidate generation. This proposed method give better performance for pattern mining. It useful to extracts the features of the words or sentences very accurately.

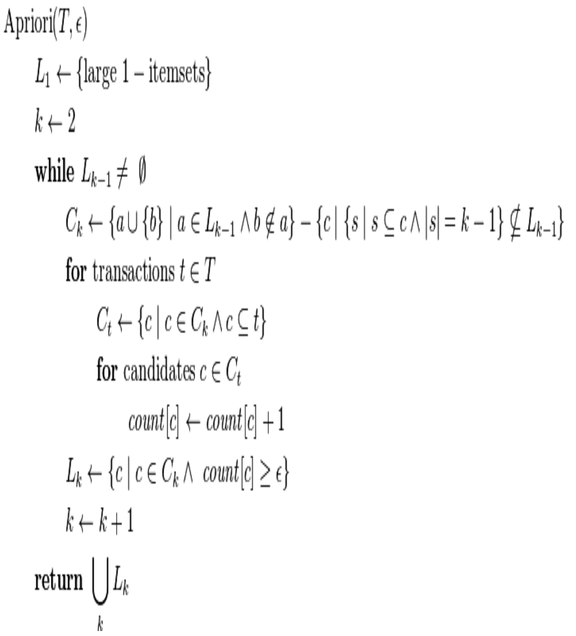
Apriori Algorithm:

Most of the parallel ARM algorithms are based on parallelization of Apriori that iterativelygenerates and tests candidate itemsets from length 1 to length k until no more frequentitemsets are found. These algorithms can be categorized into Count Distribution, DataDistribution and Candidate Distribution methods [AS96, HKK00]. This algorithm find any association rules in that data.The Count Distributionmethod follows a data-parallel strategy and statically partitions the database into horizontalpartitions that are independently scanned for the local counts of all candidate itemsets oneach process. At the end of each iteration, the local counts will be summed up across all processes into the global counts so thatfrequent itemsets can be found. The Data Distributionmethod attempts to utilize the aggregate main memory of parallel machines by partitioningboth the database and the candidate itemsets. Since each candidate itemset is counted byonly one process, and all processes have to exchange database partitions during each iterationin order for each process to get the global counts of the assigned candidate itemsets. TheCandidate Distribution method also partitions candidate itemsets but selectively replicatesinstead of partition-and-exchanging the database transactions, so that each process can beproceed independently.

The Apriori-based algorithmsare the mostwidely used because of the simplicity and easy implementation. For example pattern are usually more general and be used both positive and negative documents and large support itemsets. Also the association rules canbe directly generated on the way of itemset mining, because all the subset informationis already computed when candidate itemsets are generated. These algorithms are evaluation of term weights are based on the distribution of terms in documents. To handling all the itemsets to discovering efficient accuracy of patterns in database.

By convention, Apriori assumes that items within a transaction or itemset are sorted in lexicographic order. Let the set of frequent itemsets of size kbe Fkand their candidates be Ck. Apriori first scans the database and searches for frequent itemsets of size 1 by accumulating the count for each item and collecting those that satisfy the minimum support requirement. It then iterates on the following three steps and extracts all the frequent itemsets.

1. Generate Ck+1, candidates of frequent itemsets of size k+1, from the frequent itemsets of size k.
2. Scan the database and calculate the support of each candidate of frequent itemsets.
3. Add those itemsets that satisfies the minimum support requirement to Fk+1.Apriori algorithm is given the below:



```
Apriori($T, \epsilon$)
    $L_1 \leftarrow \{\text{large } 1 - \text{itemsets}\}$
    $k \leftarrow 2$
    while $L_{k-1} \neq \emptyset$
        $C_k \leftarrow \{a \cup \{b\} \mid a \in L_{k-1} \wedge b \notin a\} - \{c \mid \{s \mid s \subseteq c \wedge |s| = k-1\} \nsubseteq L_{k-1}\}$
        for transactions $t \in T$
            $C_t \leftarrow \{c \mid c \in C_k \wedge c \subseteq t\}$
            for candidates $c \in C_t$
                $count[c] \leftarrow count[c] + 1$
        $L_k \leftarrow \{c \mid c \in C_k \wedge \; count[c] \geq \epsilon\}$
        $k \leftarrow k + 1$
    return $\bigcup_k L_k$
```

The AprioriAlgorithm : Pseudo Code

- Join Step: Ck is generated by joining $Lk{-}1$ with itself
- Prune Step: Any (k-1)-itemset that is not frequent cannot be a subset of a frequent k-itemset.
- Pseudo-code : Ck: Candidate itemset of size k
  Lk: frequent itemset of size k

L1 = {frequent items};
for (k = 1; Lk !=∅; k++) do begin
Ck+1 = candidates generated from Lk;
for each transaction t in database do
increment the count of all candidates in Ck+1 that are contained in t
Lk+1 = candidates in Ck+1 with min_support
end
return∪k Lk;

Function apriori-gen in line 3 generates Ck+1from Fk in the following two step process:

1. Join step: Generate RK+1(first step), the initial candidates of frequent itemsets of size k + 1 by taking the union of the two frequent itemsets of size k, Pkand Qkthat have the first kí1 elements in common.

Rk+1=PkUQk= {item1, item2,......,itemk, itemk'}
Pk= {item1, item2,......,itemk, itemk}
Qk= {item1, item2,......,itemk'}
Where item1<item2<……….<itemk<itemk'

2. Prune step: Check if all the itemsets of size k in Rk+1are frequent and generate Ck+1by removing those that do not pass this requirement from Rk+1. This is because any subset of size k of Ck+1that is not frequent cannot be a subset of a frequent itemset of size k+1.

Function subset in line 5 finds all the candidates of the frequent itemsets included intransaction t. Apriori, then, calculates frequency only for those candidates generated this way by scanning the database. It is evident that Apriori scans the database at most kmax+1 times when the maximum size of frequent itemsets is set at kmax.

LDA Algorithm:

Latent Dirichlet allocation (LDA) is a generative probabilistic model of a corpus. The basic idea isthat documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words.

LDA assumes the following generative process for each documentwin a corpusD:
1. ChooseN∼Poisson(ξ).
2. Chooseθ∼Dir(α).
3. For each of theNwordswn:
(a) Choose a topiczn∼Multinomial(θ).
(b) Choose a wordwnfromp(wn|zn,β), a multinomial probability conditioned on the topiczn.

Several simplifying assumptions are made in this basic model, some of which we remove in subsequent sections. First, the dimensionalitykof the Dirichlet distribution (and thus the dimensionalityof the topic variablez) is assumed known and fixed. Second, the word probabilities are parameterized by a $k \times V$ matrix β where $\beta_{ij} = p(w^j = 1 \,|\, z^i = 1)$ which for now we treat as a fixed quantitythat is to be estimated. Finally, the Poisson assumption is not critical to anything that follows andmore realistic document length distributions can be used as needed. Furthermore, note thatNisindependent of all the other data generating variables (θandz). It is thus an ancillary variable andwe will generally ignore its randomness in the subsequent development.

Ak-dimensional Dirichlet random variableθcan take values in the(k−1)-simplex (ak-vectorθlies in the(k−1)-simplex if $\theta_i \geq 0, \; \sum_{i=1}^{k} \theta_i = 1)$, and has the following probability density on thissimplex:

$$p(\theta \,|\, \alpha) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \cdots \theta_k^{\alpha_k - 1}, \qquad (1)$$

where the parameterαis ak-vector with componentsαi>0, and whereΓ(x)is the Gamma function.TheDirichlet is a convenient distribution on the simplex — it is in the exponential family, has finitedimensional sufficient statistics, and is conjugate to the multinomial distribution. Theproperties will facilitate the development of inference and parameter estimation algorithms for LDA.Given the parametersαandβ, the joint distribution of a topic mixtureθ, a set ofNtopicsz, anda set ofNwordswis given by:

$$p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta), \quad , \tag{2}$$

wherep(zn|θ)is simplyθi for the unique I such thatz^i n=1. Integrating overθandsumming overz, we obtain the marginal distribution of a document:

$$p(\mathbf{w} \mid \alpha, \beta) = \int p(\theta \mid \alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n \mid \theta) p(w_n \mid z_n, \beta) \right) d\theta. \tag{3}$$

Finally, taking the product of the marginal probabilities of single documents, we obtain the probability of a corpus:

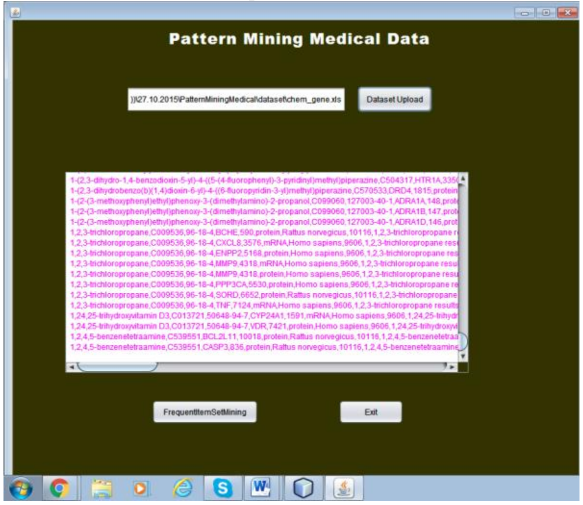
$$p(D \mid \alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d \mid \alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn} \mid \theta_d) p(w_{dn} \mid z_{dn}, \beta) \right) d\theta_d.$$

The LDA model is represented as a probabilistic graphical model. There are three levels to the LDA representation. The parametersαandβare corpuslevel parameters, assumed to be sampled once in the process of generating a corpus. The variablesθdare document-level variables, sampled once per document. Finally, the variableszdnandwdnareword-level variables and are sampled once for each word in each document.

4. EXPERIMENTAL RESULTS

The Effective pattern discovery technique for text mining has been implemented by us using Java programming language. For implementation of this system, we used java technology. The environment used for the implementation include a PC with 4GB RAM, Core 2 Dual processor. Operating system used is Windows and the IDE is Net Beans. The application form selected data set should be given in that preprocessing. This medical dataset contain genes, organism and possible diseases.

The experimental results are performed in databases from Microsoft Exel 2007. We call the algorithm based on APrioriDP and LDA. The dataset is choosing from The Comparative Toxicogenomics Database (CTD)-http://ctdbase.org/. Database results from 2004-2015 Mount Desert Island Biological Laboratory the contexts that use CTD data to the applicable CTD data pages.And the dataset CTD data name is cheme_genes. They use 11 level of parameters in the sets areGeneID, Gene name, chemicalname, chemicalID, casRN, GeneSymbol, GeneForms, organism, organismID, interaction, interactionActions, PubMedIDs. The dataset collect 1000 item data records with the size 1. The selected dataset shown in text area in the application. The application facilitates preprocessing before actual discovery of patterns. These fields are completely analyzing and discover the effective pattern to find the relevant information using the AprioriDP and LDA algorithms. The selected dataset before submitting to frequent item set proceed to configurations and compute the frequency of total itemsets. Terms with term frequency equaling to one are discarded. Fig. 1 shows that Apriori and LDA build a configurations of items in that record with size 1 and scanning the process. Next passing through the data to compute the frequency itemsetoof size and found that frequent.



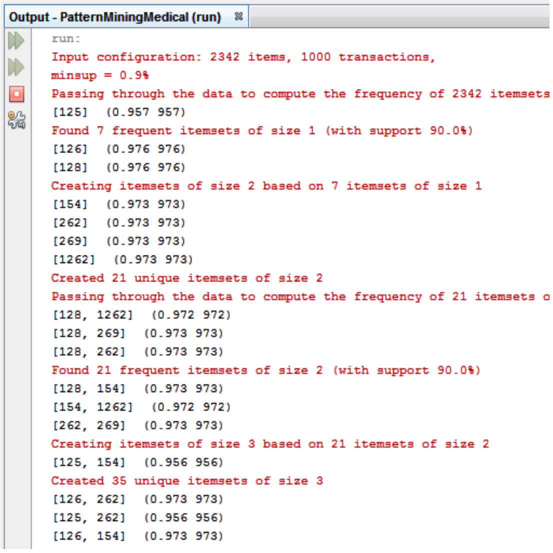
```
))\27.10.2015\PatternMiningMedical\dataset\chem_gene.xls
```

```
1-(2,3-dihydro-1,4-benzodioxin-5-yl)-4-((5-(4-fluorophenyl)-3-pyridinyl)methyl)piperazine,C504317,HTR1A,3350
1-(2,3-dihydrobenzo(b)(1,4)dioxin-6-yl)-4-((6-fluoropyridin-3-yl)methyl)piperazine,C570533,DRD4,1815,protein
1-(2-(3-methoxyphenyl)ethyl)phenoxy-3-(dimethylamino)-2-propanol,C099060,127003-40-1,ADRA1A,148,prote
1-(2-(3-methoxyphenyl)ethyl)phenoxy-3-(dimethylamino)-2-propanol,C099060,127003-40-1,ADRA1B,147,prote
1-(2-(3-methoxyphenyl)ethyl)phenoxy-3-(dimethylamino)-2-propanol,C099060,127003-40-1,ADRA1D,146,prot
1,2,3-trichloropropane,C009536,96-18-4,BCHE,590,protein,Rattus norvegicus,10116,1,2,3-trichloropropane r
1,2,3-trichloropropane,C009536,96-18-4,CXCL8,3576,mRNA,Homo sapiens,9606,1,2,3-trichloropropane res
1,2,3-trichloropropane,C009536,96-18-4,ENPP2,5168,protein,Homo sapiens,9606,1,2,3-trichloropropane res
1,2,3-trichloropropane,C009536,96-18-4,MMP9,4318,mRNA,Homo sapiens,9606,1,2,3-trichloropropane resu
1,2,3-trichloropropane,C009536,96-18-4,MMP9,4318,protein,Homo sapiens,9606,1,2,3-trichloropropane resu
1,2,3-trichloropropane,C009536,96-18-4,PPP3CA,5530,protein,Homo sapiens,9606,1,2,3-trichloropropane re
1,2,3-trichloropropane,C009536,96-18-4,SORD,6652,protein,Rattus norvegicus,10116,1,2,3-trichloropropane
1,2,3-trichloropropane,C009536,96-18-4,TNF,7124,mRNA,Homo sapiens,9606,1,2,3-trichloropropane results
1,24,25-trihydroxyvitamin D3,C013721,50648-94-7,CYP24A1,1591,mRNA,Homo sapiens,9606,1,24,25-trihydr
1,24,25-trihydroxyvitamin D3,C013721,50648-94-7,VDR,7421,protein,Homo sapiens,9606,1,24,25-trihydroxyvi
1,2,4,5-benzenetetraamine,C539551,BCL2L11,10018,protein,Rattus norvegicus,10116,1,2,4,5-benzenetetraa
1,2,4,5-benzenetetraamine,C539551,CASP3,836,protein,Rattus norvegicus,10116,1,2,4,5-benzenetetraamine
```

Fig 1. Configuration of dataset

The dataset added to the training set, and applying the support frequent item mining to evaluate the set compute and improving accuracy of frequent matching itemset. Found 7 itemset the frequent item matched and creating itemset of size to based on 7 itemsets of size 1. This result shown in fig 2. Next found another matching of itemset to create unique of itemset and passing through the dat to compute the frequency of that founding itemsets. That process continue and found again 21 frequent itemset matched on the size of 2 with supportring 90.0 % of results shown in fig2. So created itemset based on size 3To create new itemset of size 3 based on the frequent matching field in that set on the size 2 shown in the fig. 2.

We showed how the best features of the two algorithms can be combined into aAprioriDP algorithm and LDA, which then becomes the effective pattern of choice for this problem. Apriori is to count up the number of occurrences, called and support, of each data items and separate the minimal threshold matches.



```
run:
Input configuration: 2342 items, 1000 transactions,
minsup = 0.9%
Passing through the data to compute the frequency of 2342 itemsets
[125]  (0.957 957)
Found 7 frequent itemsets of size 1 (with support 90.0%)
[126]  (0.976 976)
[128]  (0.976 976)
Creating itemsets of size 2 based on 7 itemsets of size 1
[154]  (0.973 973)
[262]  (0.973 973)
[269]  (0.973 973)
[1262]  (0.973 973)
Created 21 unique itemsets of size 2
Passing through the data to compute the frequency of 21 itemsets o
[128, 1262]  (0.972 972)
[128, 269]  (0.973 973)
[128, 262]  (0.973 973)
Found 21 frequent itemsets of size 2 (with support 90.0%)
[128, 154]  (0.973 973)
[154, 1262]  (0.972 972)
[262, 269]  (0.973 973)
Creating itemsets of size 3 based on 21 itemsets of size 2
[125, 154]  (0.956 956)
Created 35 unique itemsets of size 3
[126, 262]  (0.973 973)
[125, 262]  (0.956 956)
[126, 154]  (0.973 973)
```

Fig. 2.Create unique itemsets

Compare with other Apriori and LDA to compute the support of itemset fast, so the mining time is faster. The aim of this experiment was to see how our data sets scaled with the frequent rules, independent of other factors like the database size and the number of large item sets.The Frequent item matching of the dataset capability with classification rules and similarity result over time sequences. Thisitem matches discovering the patterns in sets to mining the better results matching given the datasets. The results show that the frequent matching of terms to the corresponding discovered patterns to improve the better performance and give accurate information.

5. CONCLUSION

The effective pattern discovering over come the text mining problems. This proposed method will work for better performance in pattern mining. It extracts the features of the words or sentences very accurately. LDA is used to analyse the sentence or word with timeline features. Latent Dirichlet Allocation (LDA) based models to analyze sentence in significant variation periods, and infer possible reasons for the variations. Apriori utilizes Dynamic Programming in Frequent itemset mining. The working principle is to eliminate the candidate generation like FP-tree, but it stores support count in specialized data structure instead of tree.

It solves the following problems,

1.Same results problem for both old and new products. (The diabetes disease symptoms shows risk or not)

2.Ironic Phase problem (rumour data as actual data).

3.Problem in identification of domain in all phases

4.Sub-domain Analysis

5.Time-to-Time analysis

REFERENCES

[1] K. Aas and L. Eikvil, “Text Categorisation: A Survey,” TechnicalReport Raport NR 941, Norwegian Computing Center, 1999.

[2] R. Agrawal and R. Srikant, “Fast Algorithms for MiningAssociation Rules in Large Databases,”Proc. 20th Int’l Conf. VeryLarge Data Bases (VLDB ’94),pp. 478-499, 1994.

[3] KrissnaPriya.R., “A Improved Classification of Network Traffic using Adaptive Nearest cluster Based Classifier”, International Journal of Computer Trends and Technology ISSN:2231-2803,Vol.18, Issue No.1 January 2015.

[4] H. Ahonen, O. Heinonen, M. Klemettinen, and A.I. Verkamo,“Applying Data Mining Techniques for Descriptive PhraseExtraction in Digital Document Collections,” Proc. IEEE Int’lForum on Research and Technology Advances in Digital Libraries(ADL ’98), pp. 2-11, 1998.

[5] N. Cancedda, N. Cesa-Bianchi, A. Conconi, and C. Gentile, “Kernel Methods for Document Filtering,” TREC, trec.nist.gov/pubs/trec11/papers/kermit.ps.gz, 2002.

[6] N. Cancedda, E. Gaussier, C. Goutte, and J.-M.Renders, “WordSequenceKernels,”J. Machine Learning Research, vol. 3, pp. 1059-1082, 2003.

[7] M.F. Caropreso, S. Matwin, and F. Sebastiani, “Statistical Phrases in Automated Text Categorization,” Technical Report IEI-B4-07-2000, Instituto di Elaborazionedell’Informazione, 2000.

[8] C. Cortes and V. Vapnik, “Support-Vector Networks,”MachineLearning,vol. 20, no. 3, pp. 273-297, 1995.

[9] S.T. Dumais, “Improving the Retrieval of Information from External Sources,” Behavior Research Methods, Instruments, and Computers,vol. 23, no. 2, pp. 229-236, 1991.

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

[11] Manish Gupta, Jiawei Han, "Approaches for Pattern Discovery Using Sequential Data Mining".Applications andStudies. IGI Global, 2012.137-154. Web. 14 Nov. 2015.

[12] R. Baeza-Yates and B. Ribeiro-Neto,Modern Information Retrieval. Addison Wesley, 1999.