International Journal of Engineering & Technology, 7 (3.8) (2018) 77-81

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Mining Rare Patterns by Using Automated Threshold Support

Prof. Mangesh Ghonge[1], Miss Neha Rane[2]

[1]*Assistant Professor, Computer Department, SITRC, Nashik*
[2]*Student, M.E.(Computer), SITRC, Nashik*
* Mangesh.Ghonge@Sitrc.Org

Abstract

Essentially the most primary and crucial part of data mining is pattern mining. For acquiring important corre-lations among the information, method called itemset mining plays vital role Earlier, the notion of itemset mining was used to acquire the absolute most often occurring items in the itemset. In some situation, though having utility value less than threshold it is necessary to locate such items because they are of great use. Considering the thought of weight for each and every apparent items brings effectiveness for mining the pattern efficiently. Different mining algorithms are utilized to obtain the correlations among the information items based on frequency with the items in the dataset occurs. In frequent itemset, those things which occurs frequently whereas, in infrequent itemset the items that occur very rarely are obtained. Determining such form of data is tougher than to locate data which occurs frequently. Frequent Itemset Mining (FISM) locates large and frequent itemsets in huge data for example market baskets. Such data has two properties that are not addressed by FISM; Mixture property and projection property. Here the proposed system combines both mixture as well as projection property further providing automated support thresholds.

Keywords*: Infrequent itemset mining, minimum support, pattern mining, automated support threshold.*

1. Introduction

Infrequent pattern mining in the transaction relates to recognizing such patterns which occur rarely. This concept is adopted by frequent pattern mining which was the most focused method in data mining in early days. The pattern whose frequency of occurring is more than fixed minimum threshold frequency are called as frequent occurring pattern. Detecting the outlier from the information stream by extracting minimal frequent pattern is discussed in [1]. Detecting those rare patterns which are small in number among huge pair of data is the most challenging part. Nowadays, the thought of rare pattern mining is found in many real life application like in medical field [1][3],education[4],airports,bank, social media[6].By giving various ranges to the patterns, and by utilizing support threshold, patterns could be well classified easily. If the max threshold is defined, we are able to obtain the rare patterns or the patterns which occurs less frequently. High utility is an extension to the problem caused by frequent pattern mining. Itemsets that generates profit than min thresh-old are categorized under high itemset mining. This concept is found in market analysis, biomedicine, streamlining analysis and click stream analysis.To generate meaningful rules from low rank itemset [4][5], uses the fuzzy algorithm by generating candidate generation.It is located that some of the people behave anomalistically while surfing on net. The method to identify such people is an major issue which is solved in [6] by utilizing Sequential Topic Pattern (STP). The objective of association rule is to obtain the correlation between items. Here complexity is major challenge. It is improved by utilizing minimum support threshold. Nonetheless it at the same time frame frame allows some of the rules to help keep hidden. To fix this issue, target association rule is proposed and is discussed briefly [7], by maintaining flexibility and maintain-ing complexity Finding rare items from the transaction dataset is complicated and are hard to discover. Infrequent Weighted Itemset (IWI) is the idea where each and every item is assigned particular weight which is beneficial to mine them accord-ingly [8][9]. Detecting those rare patterns which are small in number among huge pair of data is the most challenging part. Nowadays, the thought of rare pattern mining is found in many real life application like in medical field [1][3],educa-tion[4],airports,bank, social media[6].By giving various ranges to the patterns, and by utilizing support threshold, patterns could be well classified easily. If the max threshold is defined, we are able to obtain the rare patterns or the patterns which occurs less frequently. High utility is an extension to the problem caused by frequent pattern mining. Itemsets that generates profit than min threshold are categorized under high itemset mining. This concept is found in market analysis, biomedicine, streamlining analysis and click stream analysis. To generate meaningful rules from low rank itemset [4][5], uses the fuzzy algorithm by generating candidate generation. It is located that some of the people behave anomalistically while surfing on net. The method to identify such people is an major issue which is solved in [6] by utilizing Sequential Topic Pattern (STP). The objective of association rule is to obtain the correlation between items. Here complexity is major challenge. It is improved by utilizing minimum support threshold. Nonetheless it at the same time frame frame allows some of the rules to help keep hidden. To fix this issue, target association rule is proposed and is discussed briefly [7], by main-

taining flexibility and maintaining complexity Finding rare items from the transaction dataset is complicated and are hard to discover. Infrequent Weighted Itemset (IWI) is the idea where each and every item is assigned particular weight which is beneficial to mine them accordingly [8][9].

Fig. 1: An assumed market basket (solid black circle) consist of items from two logical itemsets (red dotted circles), describing latent customer intentions.

2. Literature Survey

The outlier detection is among the important part in Data mining. Outlier may be detected by many methods which com-putes the exact distance of point in full dimension space. But because of high computation cost and curse of dimensionality the concept of proximity isn't quantitatively meaningful. [1] Discuss the outlier detection by extraction of minimal infre-quent pattern. The patterns which are obtained after mining are use whilst the descriptive of the outliers. New measures such as for example Transaction Weighting Factor (TWF), Minimal Infrequent Pattern Deviation Factor (MIPFD) and Minimum Infrequent Pattern centered on outlier Factor (MIFPDF) are employed for finding outliers. Centered on all these factors, minimum infrequent centered on outlier detection is proposed. With assistance from proposed algorithm, windowing tech-nique, proposed algorithm is found in Health care center to detect whether anyone is having healthy and or not healthy by utilizing biosensor nodes.To symbolize various classes of interesting patterns and then provide them to specific cases of rare and non present patterns, a structure of Apriori for Rare And Non present Itemset Mining (ARANIM) is proposed[2].

Result approach is apriori like and it uses the technique of traversing in bottom up incase if itemset represents set support in classical apriori approach. The resulting approach contains anti-monotone property and level wise exploration of itemset space. For this function each level which contains set of pattern is assigned by particular rank value. The pattern within particular level is use to generate pattern if next level by counting support each time and this continues until no new pattern is found. Pruning centered on anti-monotype is to optimize the moves from one level to another. Concept of utility itemset hails from frequent itemset mining.[3] proposes an algorithm using top-k algorithm for mining closed high utility itemset. The proposed algorithm computes in less time and is highly space efficient. It combines the merits of both top-k and closed high utility itemset. CHUD algorithm is mined by utilizing high utility itemset. CHUD is executed in two phases. One of the most crucial and challenging application of rare pattern mining is Education field.

It's used to find out low rank fuzzy rare itemsets for gen-erating meaningful rules from itemsets. The proposed FARIM has three phases. First phase is useful for transformation of original data into new data. Second phase is employed to generate candidate itemsets level wise. And in final phase from low rank r-itemset, fuzzy association rules are generated. Based on fuzzy Recognition Primed Decision(RPD) model, new powerful measure exclusive casual leverage is proposed

[5]. The degree of association of Casual Association Rule (CAR) may be quantified by the measures. The proposed phenomenon is found in Medical field to detect the casual relation between three drugs namely enlapril, rosuvastatin, pravastin) and ICD-9 codes. After experimenting it's clear that the proposed algorithm may be successfully make renowned Adverse Drugs Reaction (ARDs) rank among all symptoms in database. There are numerous kinds of documents streams like research paper archives, chatting messages, web form discussion etc which are made and distributed on internet. Lots of people surfing net are of abnormal behavior with susceptive nature handling these documents. In order to find and characterize such people [6] have proposed Sequential Topic Pattern (STP) and have designed the problem of mining user aware rare sequential topic pattern (URSTPs) in document stream on internet. They've completed this process in 3 phases. I Data preprocessing for topic extraction and session identifi-cation.

II STP Candidate Discovery by pattern growth by using DP based and approximation algorithm.

III User Aware Rarity Analysis by using URSTP Miner al-gorithm.

The proposed algorithm is verified by experimenting it on both real (twitter) and on synthetic database. Using the itemset tree data structure (TRAM-Rel sup), a novel targeted association mining approach is proposed in [7].

The proposed algorithm keeps the complexity(which was major dilemma of existing algorithm)is manageable by com-bining both targeted association mining query with rare rule mining Frequent association mining has complexity issue because it fails to locate all high confidence rule. Targeted rule fixes the complexity issue but is much less capable as rare rule mining. Hence proposed algorithm cover efficiency of targeted association rule and capabilities of rare rule mining. They've used Apriori algorithm alongside the synthesis of tree base approach. Lots of issue is observed while discovering rare and weighted itemsets. To overcome this problem, [8] find its solution by finding such measures that drives the Infre-quent Weighted Itemsets (IWI) mining process. Two algorithm that performs IWI and minimal IWI such as for example weighted Transaction Equivalence and Infrequent weighted Itemset Miner algorithm are proposed.The proposed algorithm results in FP Growth like algorithm that will be experimented on actual life context by domain expert resulting and validating its efficiency and effectiveness.

[10] As just like [8],[9] deals with the situation of discovering rare and weighted itemset. Here, the performance of IWI mining algorithm is improved by using FP Growth structure. The proposed algorithm reduces execution time hence improv-ing complexity issue by proposing tree based method which removes frequent occurring items by the method of pruning. Using multiple minimum support, many existing algorithms are useful for miming Frequent patterns. But many of them requires a lot of time to offer result and also uses large amount of space. This matter is taken into consideration in [10] by proposing enumeration tree ME

structure (FP-ME) with multiple support, new Sorted Downward Closure (SDC) property and Least Minimum Support (LMS) is useful for pruning of FP-ME tree. Because of pruning the problem of space is solved and predicated on FP-ME algorithm, the situation of rare item is efficiently solved.

In [11],the have discussed basic concepts of rare patterns and various types and algorithms for rare patterns like Apriori algorithm, Apriori inverse algorithm, Relative Support apriori, Frequent Pattern growth algorithm, tree based approach etc. Also they have discussed about areas in which rare pattern plays important role.[17] has put forward the important point that till now, all the research has been done on rare sequential pattern is cramped in static field only. Hence they have proposed rare sequential pattern over data stream using sliding window and also have proved it efficient

For mining rare pattern ,one of the concept called nave approach is also used.[24]. They have followed two steps to come to final results. One is to split data into frequent itemsets and rare itemset mining. For this purpose they have used nave based algorithm and optimized one.[32] has introduced again high utility infrequent itemsets using Utility Pattern Rare Itemset algorithm. They have proposed an approach which is based on dynamic approach for high utility itemset.

3. System Architecture/ System Overview

Problem Statement: In some situations, e.g when the requirement is always to minimize a specific cost function, discovering rare data correlations is more interesting than mining frequent ones[15].[7] considers the two drawbacks of traditional sequential pattern ,i.e consideration of only frequent sequence and other is restricted to the static environment of data set only. This study emphasizes an approach to obtain the infrequent itemsets involving rare items by setting the support thresholds automatically by using logical itemset mining.

4. System Analysis

4.1. Proposed System

The proposed method combines Apri-ori and MS-Apriori to mine logically rare itemsets among huge amount of data. Three different groups namely Most interesting group(MIG), Rare interesting group(RIG) and some what interesting groups (SIG)are minned to obtain rare item-sets.Among all these groups, Calculated levelwise automated support thresholds like apriori is used by most interesting and rare interesting groups. Whereas, rare interesting group uses support thresholds like MS Apriori i.e. Automated Apriori Rare(AAR). When all type of data is gathered at one place such place is called as dataset.By applying the technique of preprocessing,raw data is ommited and hence pure data is obtained called set of transactions.After unique items are identified from these transaction,support count is calculated by using the standard formula. High occurrence noise is removed and is further processed for preserving low occurrence signal. From this, only rare items are inserted after finding rare transaction. Hence MRCP tree is generated and as a result, set of rare items are obtained.MRCP tree sacns the data once which reduces the space as well as the time complexity of the system. Hence by using the hybrid approach of Apriori, MS Apriori and Logical itemset mining,the proposed system efficiently mine all the rare items from huge data.The system will cover all the drawbacks of Fp Growth algorithm,sliding window,sequential algorithm.etc.

One most important benifit of this system is that it will decide the value of threshold on its own and hence the result obtained will be accurate accordingly.

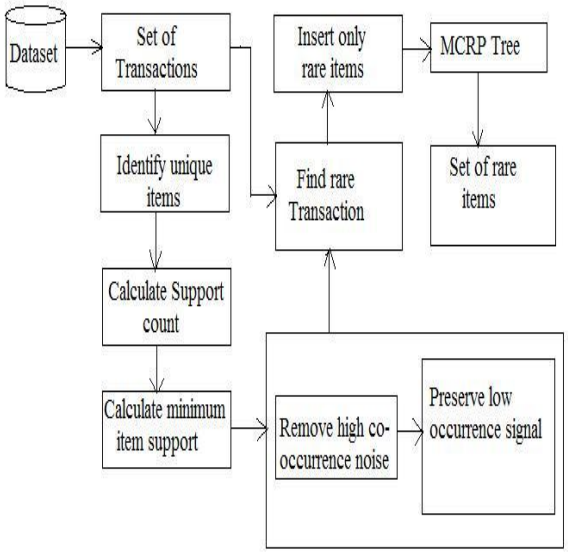

Fig. 2: Block Diagram of Proposed System

4.2. System Design

A system design contains a usecase diagram which specifies admin and user in diagramatic representation of proposed system.In shows the relation and how the admin and user are related to each other and their assignments.

The admin and user are assigned their task according to the assignments between them and usecases.

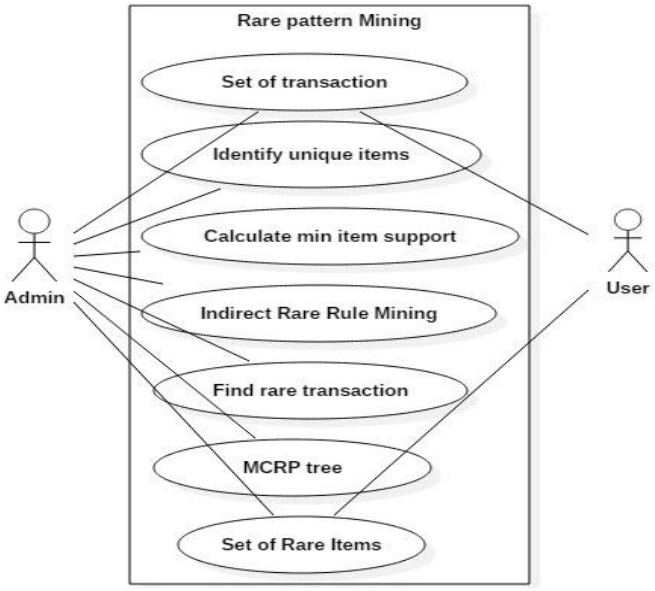

Fig. 3: Flow of system

4.3. Mathematical Model

Let system S be defined as

$S = \{T, U, R, MIS, SUP, H_C, L_C\}$

T = set of transactions

$T=\{t_1,t_2,t_3,\ldots\ldots,t_n\}$
U = set of unique items
$U=\{ U_1,U_2,U_3,\ldots..U_n\}$
R= set of rare items
$R =\{r_1,r_2,r_3,\ldots..,r_n\}$

MIS = set of minimum item support value
$MIS=\{m_1,m_2,m_3,\ldots,m_n\}$

N=total no of items

SUP= set of support count for each item.
$SUP=\{s_1,s_2,s_3,\ldots..s_n\}$
H_C = High co-occurrence noise
L_C =Low co-occurrence noise

f1 → function f1 read set of transaction as input as find set of unique items.
$f_1(T) \rightarrow [(t_1,t_2,\ldots t_n) \rightarrow (U_1,U_2,\ldots U_n)] \in U$

f_2 → function f_2 read set of unique items is input and find support count and minimum support count
$f_2(U) \rightarrow [(U_1,U_2,\ldots U_n) \rightarrow [(s_1,s_2,\ldots s_n)(m_1,m_2,\ldots m_n)] \in (MIS,SUP)]$

f_3→function f_3 read item with minimum support count and remove high co-occurrence pattern and presence low co-occurrence signals
$f_3(MIS) \rightarrow \{(m_1,m_2,\ldots m_n) \rightarrow H_C) \subseteq (T_t)\} \in H_C$
where T_t = total MIS set count

f_4→ function f_4 reads H_C pattern and find rare transactions
$f_4(H_C) \rightarrow \{(h_{c1},h_{c2},\ldots h_{ck}) \rightarrow (r_1,r_2,\ldots r_p)\} \in R$
Where P<<k
P=total count of rare itemset

f_5→Function f_5 read rare itemset and creates MCRP tree for fast access.
$f_5(R) \rightarrow$ {MCRP}

5. Performance Analysis

The size and capacity requirements are also important.Our system can be efficiently run on Pentium IV system with minimum 512 MB RAM. For calculating experimental result following dataset were used,
1.Trasaction dataset of chess is taken as first field from which rare moves are classified and sorted.
2.Transaction dataset of food mart is taken as second field from which rare combinations of foods are classified.
For classifying the patterns,first the concept of high co-occurrence noise is removed and then rare items are classified using Normalized Point-wise Mutual Information.

LISM-Consistency
FISM depends on the support i.e. frequency as a vital statistic on itemsets.LISM is just like finding all frequent itemsets of size 2 with an assistance threshold of co-occurrence

High Co-occurrence Noise
A pair of common products such as for example DVD and Shoes sold with a retailer. Since both are high volume by themselves, they might co-occur in a sizable amount of market baskets

Low Co-occurrence Signal
A pair of rare products such as for example home-theatre-system and high-definition-TV. To keep such low frequency co-occurrences, the support threshold must be reduced substantially, which consequently will result in addition of lot of spurious product pairs.

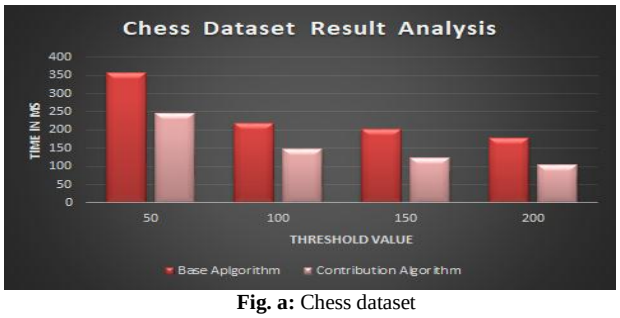

Fig. a: Chess dataset

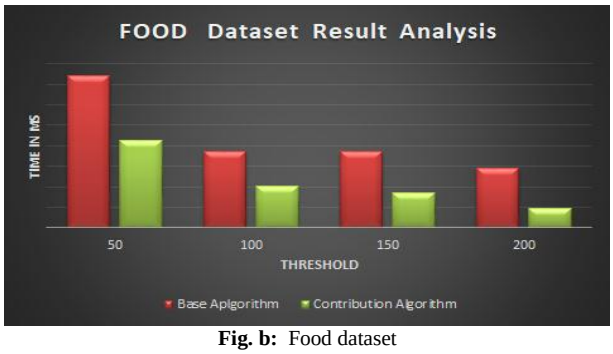

Fig. b: Food dataset

In fig a and fig b,transaction set of chessand fpood is considered respectively which is drawn time against threshold.It can be seen that the time required by contributed algorithm is less as compared to base algorithm in both cases

6. Conclusion

The key advantage to perform Infrequent itemset mining was to enhance the profit of rarely found datasets in the transactions. The very first attempt is to find the frequent item set mining and then to reveal the infrequent weighted item sets. There are many existing algorithms from sequential pattern growth to window sliding but the sets of candidate generation is huge. The proposed system uses hybrid approach by using both Apriori and MS Apriori techniques for mining logical itemsets. For mining infrequent patterns and which becomes the cornerstone for future years work likely to be achieved in this area of pattern mining.

Acknowledgement

I would sincerely like to thank our Professor Mangesh Ghonge, Department of Computer Engineering, SITRC, Nashik for his guidance, encouragement and the interest shown in this project by timely suggestions in this work. His expert suggestions and scholarly feedback had greatly enhanced the effectiveness of this work.

References

[1] C. Sweetlin Hemalatha, V. Vaidehi,and R. Lakshmi, "Minimal infer quent pattern based approach for mining outliers in data streams", Jour nal on Expert Systems with Applications, Elsevier , 2014.
[2] Mehdi Adda , Lei Wu , Sharon White, and Yi Feng, "Pattern De tection withRare Itemset Mining, International Journal on Soft C puting, Artificial Intelligence and Applications (IJSCAI), Vol.1, No.1, August 2012.
[3] Anu Augustin,Vince Paul and Vishnu G. Nair, "High Utility Itemset Mining withTop-k CHUD (TCHUD) Algorithm", International Journal of Computer Applications, 3 May 2017
[4] Cheng-HsiungWeng, High Utility Itemset Mining withTop-k CHUD (TCHUD) Algorithm, Elsevier Journal 13 February 2011.
[5] Hao Ying, John Tran, Peter Dews,Ayman Mansour, and R. Michael Massanari, "A Method for Mining Infrequent Causal Associations and Its Application in Finding Adverse Drug Reaction Signal Pairs", IEEE Transaction,4 April,2013

[6] Jiaqi Zhu, Yunkun Wu, Zhongyi Hu, and Hongan Wang, "WangMining User-Aware Rare Sequential Topic Patterns in Document Streams", IEEE Transaction,2016.
[7] Jennifer Lavergne, Ryan Benton and Vijay V. Raghavan , "TRARM-RelSup: Targeted Rare Association Rule Mining Using Itemset Trees and the Relative Support Measure", Springer, 2012.
[8] C. Luca Cagliero and Paolo Garza, "Infrequent Weighted Itemset Mining Using Frequent Pattern Growth", IEEE Transaction on Knowledge and Data Engineering ,4, APRIL 2014 Busan, Korea.
[9] A. Jalpa A Varsur1, Nikul G Virpariya , "Mining Rare Itemset Based on FP Growth Algorithm", International Conference
[10] Wensheng Gan, Jerry Chun-Wei Lina, Philippe Fournier-Viger, Han-Chieh Chaoa,c, Justin Zhan "Mining of frequent patterns with multiple minimum supports",Elsevier,2017
[11] Yun Sing Koh and Sri Devi Ravana, "Unsupervised Rare Pattern Mining: A Survey", ACM Transactions on Knowledge Discovery from Data, 2016.
[12] Saeed Piri, Dursun Delen, Tieming Liu, William Paiva, Development of a New Metric to Identify Rare Patterns in Association Analysis:The Case of Analyzing Diabetes tions ,10.1016/j.eswa.2017.09.061
[13] Timothy M. Hospedales, Shaogang Gong, and Tao Xiang, "Finding Rare Classes: Active Learning with Generative and Discriminative Models", IEEE Transaction, 2013.
[14] Jayakrushna Sahoo1Ashok Kumar Das, A. Goswami1, "An effi-cient fast algorithm for discovering closed high utility itemsets"
[15] Ashish Gupta, Akshay Mittal, Arnab Bhattacharya, "Minimally Infre-quent Itemset Mining using Pattern-Growth Paradigm and Residual Trees, 17th International Conference on Management of Data ,2011
[16] Varsur Jalpa A.,Desai Sonali P., Hathi Karishma B, "Performance Analysis of Rare Itemset Mining Algorithms",Journal of Emerging Technologies and Innovative Research (JETIR),2015.
[17] Weimin Ouyang, Mining Rare Sequential Patterns in Data Streams with a Sliding Window ,The 2016 3rd International Conference on Systems and Informatics (ICSAI 2016).
[18] Fernando Benites, Elena sapozhnikova, "Evaluation of Hierarchical Interestingness measures for mining pairwise generalized association rules", IEEE Transaction, 2014
[19] Kantarcioglu, Chris Clifton, "Privacy Preserving distributed Mining of Association Rules on Horizontally partitioned Data",IEEE Transaction, 2004
[20] Thiago Henrique Cupertino, Murillo Guimares Carneiro, Qiusheng Zheng, Junbao Zhang, Liang Zhao, "A Scheme for High Level Data Classification Using Random Walk and Network Measures,2015. 10.1016/j.eswa.2017.09.014
[21] Sheethal Abraham, Sumy Joseph, "Rare And Frequent Weighted Itemset Optimization Using Homologous Transactions: A Rule Mining Ap-proach", J, 2015 International Conference on Control, Communication and Computing India (ICCC) ,November 2015
[22] Jerry Chun, wensheng Gan, Philippe Fournier, "High Utility mining and Privacy prteserving utility mining, Elsevier, 2016
[23] Tamir Tassa, "Secure Mining of association Rules in Horizontally Distributed Database", IEEE Transaction, 2013
[24] Luca Cagliero, Discovering Temporal Change Patterns in the Presence of Taxonomies, IEEE Transaction, 2013
[25] Sarra Gacem, Djamila Mokeddem , Hafida Belbachir, "Privacy Preserv-ing In Data Mining: Case of Association Rule", IJCSI, 2013
[26] Shen Zhong, "privacy preserving algorithms for Distributed mining of frequent itemsets", Elsevier, 2007
[27] Luigi Troiano, Giacomo Scibelli, Cosimo Birtolo , "A Fast Algorithm for Mining Rare Itemsets", 2009 Ninth International Conference on Intelligent Systems Design and Applications.
[28] A. Nor Antonina, N. A. M. Shazili, B. Y. Kamaruzzaman, M. C. Ong, Y. Rosnan, F. N. Sharifah "Geochemistry of the Rare Earth El ments(REE) Distribution in Terengganu Coastal Wa ters: A Study Case from Redang Island Marine Sediment ", 2013 http://dx.doi.org/10.4236/ojms.2013.33017
[29] Mehdi Adda1, Lei Wu2, Yi Feng3 , " Rare Itemset Mining", Sixth International Conference on Machine Learning and Applictions,2007.
[30] Monika Akbar, Rafal A. Angryk "Frequent Pattern-Growth Ap proach for Document Organization",ONISW, 2008
[31] Junfeng Ding, Stephen S.T. Yau "TCOM, an innovative data struc ture for mining association rules among infrequent items",Elsevier 2009.
[32] Laszlo Szathmary, Petko Valtchev "Towards Rare Itemset Mining",19th IEEE International Conference on Tools with Artificial Intelligence,2007.
[33]PaoloGarza,FabioPulvirenti,LucaVenturin "Frequent Itemsets Minin for Big Data: A Comparative Analysis",https://doi.org/10.1016/j.bdr.2017.06.006
[34]Ms.KalyaniTukaramBhandwalkar,Ms.MansiBhonsle "Study of Infrequent itemset mining Techniques",International Journal of Engineering Research and General Science Volume 2, Issue 6, October-November, 2014.
[35]Ashish Gupta, Akshay Mittal, Arnab Bhattacharya "Minimally Infrequent Itemset Mining using Pattern-Growth Paradigm and Residual Trees",17th International Conference on Management of Data ,2011