ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 2, Issue 2, pp: (53-59), Month: April-June 2014, Available at: www.researchpublish.com

DISCOVERY OF FREQUENT AND NON-REDUNDENT ITEMSET USING HIGEN MINER

S.KIRUTHIKA[1], T.SHEIK YOUSUF[2]

[1]PG Scholar, Department of Computer Science and Engineering
[2]Associate professor, Department of Computer Science and Engineering
Mohamed Sathak Engineering College, Kilakarai.

Abstract: **Data mining is important and useful tool amongst several systematic tools for extracting data that are accumulated in a database. Frequent pattern mining is one of data mining, which will extract the items that occur more number of times. Data may be of any kind, facts, numbers, or text. In recent years, Organizations are accumulating huge and increasing amounts of data in dissimilar formats and different databases. Frequent itemset mining, which focuses on finding a relationship among data. Change mining, detects and report any considerable changes if occurs in the set of mined itemsets from different time periods. This project extends the vibrant change mining problem, in the framework of frequent itemsets, by exploiting recurrent generalized itemsets to characterize information linked with infrequent patterns. To address this issue, I introduce two novel kinds of vibrant patterns, namely the HIGEN MINER (History Generalized pattern) and NON-REDUNDANT HIGEN MINERs. HIGEN MINER detects the mined frequent itemset, if any items become infrequent during next extraction those changes are focused and avoided by Apriori based support driven approach. NON-REDUNDANT HIGEN MINER, which stores items whose support value is minimum, these infrequent items may become frequent during next extraction. If it remains same, then these infrequent items are discarded.**

Keywords: **Frequent itemset mining, Change mining, Association Rule Extraction, Higen Miner, Non-Redundent Higen Miner.**

I. INTRODUCTION

In general, Frequent itemset mining which extracts the items that are frequently occurring by calculating support threshold value. If support value is minimum, the item has to be entered into infrequent itemset. If support value is maximum or equal then it has to be considered as frequent items.

Frequent pattern mining is mainly used in Market analysis, medical image processing. In super market, if a product is combinely bought with some other product continuously. They will mark it as frequent items. The marketers will start to sell the product together until next match appear. For example, if Soap and lufah pad are brought together based on support count it is marked as frequent items.

Consider two itemset I1 and I2, two consecutive months are compared on the basis of any one of the date, time, location, item description and frequent itemset are extracted by support threshold value. The example is given below in Table 1.1 and Table 1.2.

The result itemset contain frequent items and also infrequent items i.e., items with low support value. Note below table under proposed system shows items with low support value. These may be considered at time of monitoring. A dataset is a collection of records which contain all items. Using HIGEN MINER frequent items are extracted and Infrequent items are place separately in NON-REDUNDENT HIGEN MINERs which may be future revised to check whether infrequent itemset become frequent items or not.

II. RELATED WORK

1) Generalized Itemset Discovery

The rules that are generated by support and confidence are difficult to analyse which is called Association rule extraction, even if their buried data might be relevant.

To analyse similarities between data's are done by powerful and effective tools where even some buried information are extracted by previous approaches. The objective is to monitor tactic that balance the data immoderation load and better utilize a multi-core cluster system for data mining application. The main issue in this paper is low performance.

It won't mine all frequent generalizations of an atypical pattern, but slightly generates only the one characterized by low redundancy.

2) Change Detection in Datasets

This paper presents a sketch out to detects changes inside a data set at imaginary level very deeply. The main idea is to obtain a rule-based illustration of the data set at different time period and to rarely analyse how these regulations change.

Discovering changes and acting upon to or them or before any other itemset others has become a tenaciously issue for many organisation. Existing data analysis techniques shows that task under consideration is stable over number of times due to assumption. Here detection and changing are made at imaginary level only it may not be real or true.

The main drawback is discussed previously. Imaginary may sometimes workout but not all the time. The detection based on real dataset should be made to get accurate result.

3) Frequent Generalized Pattern

FGP(Frequent Generalized Pattern) takes input as hierarchy. Then categories it and produces generalized itemset by using association rules which happens same in generalized itemset. The results generated by using association rules are strongly recommended by users not automatically generated.

The FGP algorithm modifies and combines result produced by two algorithm. The result contains unordered itemsets. This algorithm extract itemset by processing transactional database.

FGP+

The drawback in FGP is recovered by FGP+. It is proficient and solves numerous problems of pattern extraction, such as the expensive creation of Training date set sets and the over-generalization of rules. The ways to achieve parallelism and in turn abridged computational time by employing task. Item set parallelism in multiprocessing and multi-computing situation is analysed in this paper. Another important drawback is DPARM load balancing, which is also recovered. But the drawback in FGP+ is dynamic load balancing where both candidate set and computation task are handled.

4) Cubegrades

Cubegrades are expensive compared to other tools. These are represented to set measurement which is affected by modifying a cube through specialization. Thus the name cubegrade arose. Hence roll up, roll down, mutation process takes place. Roll up also called generalization, roll down also called specialization and finally mutation is called aggregates. Mutations changes based on cubes dimension.

In addition to these an important concept used in cubegrades is "COUNT" which is to measure and capture trends. In this paper also association rule is used to examine itemsets. Cubegrades are atoms which supports difficult "what if" scrutiny tasks handling with behaviour of subjective aggregates over various database segments. Cubegrades applications are marketing, sales, banking, business and other data mining application.

The association rule is viewed as description about how the cube represents the rules and how it checks the itemset which is affected by specializing it by adding an extra constraint expressed by adding some functionality. The confidence of association rule is viewed as the share of the support value, when the cube analogous to the remains of a rule.

The drawback in cubegrades which is discovered later. Cubegrades is more expensive, using of COUNT is useless because cubegrades use other typical aggregate procedures in addition to COUNT.

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 2, Issue 2, pp: (53-59), Month: April-June 2014, Available at: www.researchpublish.com

5) Rule discovery

Rule Discovery is important technique were projected to help the user to find appealing rules. Data reduction is important concept in rule discovery. This reduces data into small fragments and evaluate approximately. Initially this paper idea is get result accurately but later it does not provide correct result due to fine fragmentation.

Later these techniques employ the whole dataset and extracts completely and then filter on rank basis in many ways. This was also not a sufficient method.

The critical part is it adapts and modifies to changes easily. The buried data's can be revealed and given to normal users in paragraph format. Predicts any changes if happens. The main drawback fragmenting then combining may produce inaccurate result sometimes.

Here performance evaluation happens by evaluating synthetic data's. This makes the self of real world entities which are represented in the data. Fault-tolerance is another drawback. Here, the computational speed and purpose is not required. Alternative to Fault-tolerance may used occasionally but not every time.

New technique has been formed to answer the question by experts. The problem was well analysed and came to a decision that rule interesting which will allow to select a better rule is the best solution. Though this technique is expensive but very useful to correct problems.

6) A Fuzzy Approach

Generally fuzzy means real world entities which is unclear about the way it performs. To induce the performance linguistic variables and linguistic terms were used to represent the discovered association rules. Particular, fuzzy decision trees are structured to determine the changes in the revealed association rules. The unclear decision trees are then transformed to a group of fuzzy rules, called fuzzy meta-rules. Meta rule in general called as are rules about rules. By performing this, the changes veiled in the facts can be discovered and accessible to normal users If any changes persist then it has to be entered into itemset report.

Generally changes happens frequently, nothing is constant in life. Thus the characteristic of fuzzy approach is handled in many ways. Each data record contains many data's, each data should be partitioned and each contains a set of granulated data which are collected in different time period. The drawback in this approach is it doesn't reveal certain data's during vast difference time period.

III. PROPOSED SYSTEM

In proposed concept, Apriori based support driven approach is used. HIGEN is used to measure the frequent item set in given data set. The HIGEN MINER algorithm is used to find frequent itemset in particular time interval. Here multiple generation of itemset over different time period is avoided. Generally data's are scattered all over internet and stored in various forms or may contain dissimilar or missing entities.

Data cleaning which removes noisy, irrelevant data and not just about removing bad data or missing values, but also discovering hidden correlations in the data, finding where the data came from that are most accurate, and influential which parts are the majority appropriate for use in scrutiny.

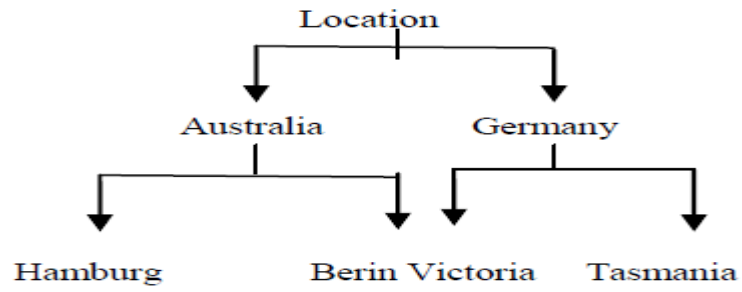

Fig 1.1: Taxonomy

Fig 1.1 shows the taxonomy by which the product's location is classified. For example, Location where the product exported or imported in consecutive months are divided in two place. Those places are further divide based on

product's comparly bought together. The data shows if a customer bought product before the product was obtainable on the market, or that customer shops frequently at a store located some miles away from their home.

The possible combinations generation over product is avoided, similarly to, an Apriori-based support-driven generalized itemset mining approach, in which the generality procedure is triggered on infrequent itemsets only. Same as the generalization process does not generate all possible ancestors of an infrequent itemset at any hierarical level, but it terminate at the generalization level in which at least a frequent ancestor produced.

Multiple taxonomy evaluations over the same pattern, the generalizing procedure of each itemset is belated after its support estimate in all time stamped itemsets and hierariclay applied on infrequent indiscriminate itemsets of escalating generalization level.

To reduce the extraction time, HIGEN generation is performed quickly, without the need of a post processing step. Also, a customized adaptation of the HIGEN MINER algorithm is anticipated to address NON-REDUNDANT HIGEN extraction.

HIGEN

Initially itemset contain both frequent items and infrequent items in an ordered sequence. The support value is calculated using,

$$\text{Support} = \frac{|\text{cov}(A,I)|}{|I|}$$

I= Itemset where all items like frequent and infrequent items are present
A= it is a set of records

HIGEN generate all possible combination and will reject infrequent items whose support value does not match. This is a drawback. Infrequent items may become frequent after sometimes or during next comparison. Rejection of those infrequent items will reduce exactness.

HIGEN MINER

HIGEN MINER address the problem in HIGEN. It stores the infrequent items for later consideration, which will not happen in HIGEN where the infrequent items are rejected initially and will never come into consideration.

HIGEN MINER will automatically select the items from two consecutive months. Based on support value it will divide all items into frequent itemset and infrequent itemset.

Both frequent and non frequent items are kept and should be monitored to check any changes occurred. If any changes occurred during next monitoring the changes should be corrected.

The changes may be frequent items become infrequent or infrequent items become frequent at the time of monitoring then it should be changed.

NON-REDUNDENT HIGEN MINERs

Among the set of mined itemsets, the frequent items are placed in HIGEN MINER and infrequent items are placed under NON-REDUNDENT HIGEN MINER. The infrequent items will be monitored over certain time period to take changes into account if any.

HIGEN Categorization and Selection

Province expert are generally in lay the responsibility of looking into the discovered temporal change patterns. To validate the discovered temporal change patterns, I am looking in accusation to emphasize most prominent trends.

HIGENs are classified based on their instance related tendency, they are 1) Stable HIGENs, i.e., HIGENs that are generalized itemsets which belongs to the similar generality level, 2) monotonous HIGENs, i.e., HIGENs that are in order of monotonous trend, and 3) oscillatory HIGENs, these HIGEN are also sequential but shows variables and non-monotonous tendency.

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 2, Issue 2, pp: (53-59), Month: April-June 2014, Available at: www.researchpublish.com

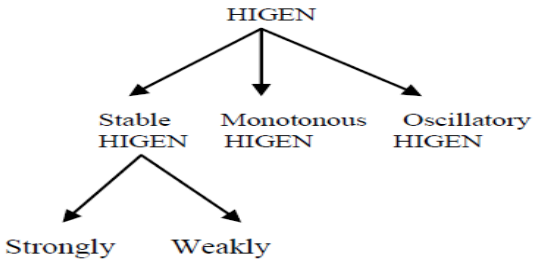

Fig 1.2: HIGEN Classification

Fig 1.2 shows the classification order of HIGEN through which categorization are made.To shrink the quantity of generated itemsets, organizers should focus on the set of NON-REDUNDANT HIGENs, where only minimal support value items are categorized. By checking NON- REDUNDANT HIGENs we can be able to notice the items which are infrequent became frequent or not after some time period.

Note mining of the NONREDUNDANT HIGENs may be proficient by vaguely altering the HIGEN MINER. The HIGEN MINER and should be frequently updated to know whether any changes occurred.

Table 1a: Items in March

Time	**Date**	**Location**	**Product**
10.00am	02-03-2014	Victoria	Shorts
12.00pm	05-03-2014	Hamburg	Night dress
3.00pm	06-03-2014	Victoria	Shorts
4.00pm	08-03-2014	Berin	Shorts

Table 1b: Items in April

Time	**Date**	**Location**	**Product**
10.15am	03-04-2014	Berin	Night dress
12.40pm	06-04-2014	Hamburg	Night dress
3.03pm	08-04-2014	Victoria	Shorts
4.05pm	10-04-2014	Berin	Shorts

Table 2: Extracted itemset

HIGEN from I1 to I2	
{Shorts} {sup=2}	→{shorts}{sup=1}
{Night dress}{sup=1}	→{Night dress}{sup=2}
{Shorts,berin}{sup=2}	→{Shorts,victoria}{sup=2}
{Nightdress,Hamburg}{sup=2}	→{Nightdress,Berin}{sup=2}

Fig 1.3 given below shows how HIGEN MINER is enhanced. First all data's are collected and processed using HIGEN. After that frequent and infrequent items are stored separately using HIGEN MINER. HIGEN MINER will

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 2, Issue 2, pp: (53-59), Month: April-June 2014, Available at: www.researchpublish.com

monitor for changes, if any changes happens then it should be marked and changes will be made.Finally all frequent items are extracted.

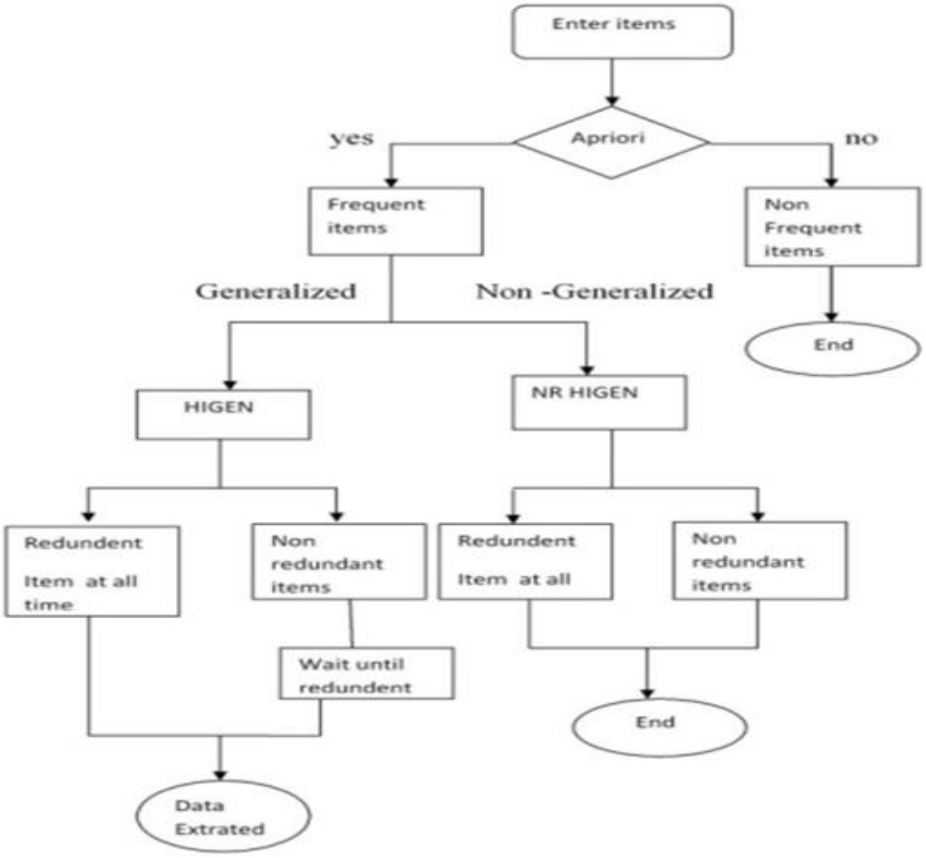

Fig 1.3: Flow Diagram

IV. CONCLUSION AND FUTURE WORK

This paper address the problem of change mining in the perspective of frequent itemsets, two vibrant patterns, namely the HIGEN MINER and the NONREDUNDANT HIGEN MINER. This addresses the problem in HIGEN mining by means of a post dispensation step after performing the traditional generalized itemset mining step.

The evolution of itemsets in various time periods without pruning relevant but uncommon knowledge due to minimum support threshold, to extract generalized itemsets that are characterized by negligible redundancy if sometimes one itemset become infrequent after sometimes then it should be marked as infrequent. The usefulness of the HIGEN MINER is mainly used by super market or Import –Export companies or showrooms which calculate the similarities. This paper will reduce burden of owners thus by making calculations easy.

Future work will address: 1) the repeated presumption and use of multiple hierarchical form of data from real world (e.g., public network), and 2) More complex HIGEN are pushed to make quality constraints into the extraction process.

REFERENCES

[1] Luca Cagliero, "Discovering Temporal change patterns in the presence of Taxonomies",IEEE transaction on knowledge and Data Engineering, vol.25,no 3March 2013.

[2] E. Baralis, L. Cagliero, T. Cerquitelli, V. D'Elia, and P. Garza, "Support Driven Opportunistic Aggregation for Generalized Itemset Extraction," Proc. IEEE Fifth Int'l Conf. Intelligent Systems (IS '10), 2010.

[3] M.Bo¨ttcher, D. Nauck, D. Ruta, and M. Spott, "Towards a Framework for Change Detection in Datasets", Research and Development in Intelligent Systems XXIII, M. Bramer, F. Coenen, and A. Tuson, eds., pp. 115-128, Springer, 2007.

[4] G. Dong, J. Han, J.M.W. Lam, J. Pei, K. Wang, and W. Zou, "Mining Constrained Gradients in Large Databases", IEEE Trans. Knowledge and Data Eng., vol. 16, no. 8, pp. 922-938, Aug. 2004.

[5] G.Dong and J. Li, "Mining Border Descriptions of Emerging Patterns from Dataset Pairs",Knowledge and Information Systems, vol. 8, pp. 178-202, Aug. 2005.

[6] W.-H. Au and K.C.C. Chan, "Mining Changes in Association Rules: A Fuzzy Approach,"Fuzzy Sets Systems,vol. 149, pp. 87-104, Jan. 2005.

[7] S.C. Gates, W. Teiken, and K.-S.F. Cheng, "Taxonomies by the Numbers: Building High-Performance Taxonomies,"Proc. 14th ACM Int'l Conf. Information and Knowledge Management (CIKM '05), pp. 568-577, 2005.

[8] P. Giannikopoulos, I. Varlamis, and M. Eirinaki, "Mining Frequent Generalized Patterns for Web Personalization in the Presence of Taxonomies,"Int'l J. Data Warehousing and Mining, vol. 6, no. 1, pp. 58-76, 2010.

[9] T. Imieliski, L. Khachiyan, and A. Abdulghani, "Cubegrades: Generalizing Association Rules,"Data Mining and Knowledge Discovery,vol. 6, pp. 219-257, 2002, doi: 10.1023/A:1015417610840.

[10] B. Liu, W. Hsu, and Y. Ma, "Discovering the set of Fundamental Rule Changes,"Proc. Seventh ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining,pp. 335-340, 2001.

[11]B. Liu, Y. Ma, and R. Lee, "Analyzing the Interestingness of Association Rules from the Temporal Dimension,"Proc. IEEE Int'l Conf. Data Mining (ICDM),pp. 377-384, 2001.

[12] L.D. Raedt, "Constraint-Based Pattern Set Mining,"Proc. SIA Int'l Conf. Data Mining,pp. 237-248, 2007.

[13]B. Shen,M.Yao,Z. Wu, and Y.Gao,"Mining Dynamic Association Rules with Comments,"Knowledge and Information Systems,vol. 23, pp. 73-98, Apr. 2010.

[14] P.-N. Tan, V. Kumar, and J. Srivastava, "Selecting the Right Interestingness Measure for Association Patterns," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '02), pp. 32-41, July 2002.

[15]Y. Tao and M.T. O'zsu, "Mining Frequent Itemsets in Time-Varying Data Streams",Proc. 18th ACM Conf. Information and Knowledge Management (CIKM '09),pp. 1521-1524, 2009.