ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

Simplified Data Search on Frequent Weighted Item-set Mining by means of Map Reduce

Anusha M, Dr. Chetana Prakash.

M.Tech Student, Dept. of Computer Science and Engineering, Bapuji Institute of Engineering and Technology, Davangere, India.

PhD, HOD in MCA, Dept. of Master of Computer Application, Bapuji Institute of Engineering and Technology, Davangere, India.

ABSTRACT*:* This paper describes Simplified data search on frequent weighted item-set mining by means of map reduce, mining is a variation of frequent item set mining where it finds the recurrent patterns i.e., it finds the data items which occur very oftenly. Here the data which is stored in HDFS directory is encrypted because those data are secured base. And considering weight for each distinct item in a transaction independent manner adds effectiveness for finding frequent item set mining. The acquired information is expansive, heterogeneous, and created at fast, in this manner it gets to be hard to adapt. There are different frequent items mining algorithms are available, however, there is no parallel and circulated solution for mining weighted regular items from huge information has never been suggested. Several articles related to frequent and weighted infrequent item set mining were proposed. This paper focus on reviewing various Existing Algorithms related to frequent and infrequent item set mining which creates a path for future researches in the field of Association Rule Mining.Frequent item set mining is one of the popular data mining techniques and it can be used in many data mining fields for finding highly correlated item sets. Infrequent item set mining finds rarely occurring item sets in the database. Most of the Existing Infrequent item set mining techniques finds infrequent weighted item sets with high computing time and are less scalable when the database size increases.

KEYWORDS: Infrequent weighted item-set, Frequent pattern growth, Data Mining, Frequent pattern Mining, Weighted mining

I. INTRODUCTION

Data mining is a process of discovering interesting patterns, such as itemsets, sub sequences, associations, or classifiers, where interestingness measures play an important role. With frequent itemset mining, an itemset is regarded as interesting if its occurrence frequency exceeds a user-specified threshold. Frequent itemset mining has been a research area for decades with tremendous progress having been made, among which are Apriori and FP-growth. Both Apriori and FP-growth employ an anti-monotone property to prune search space: A superset of an infrequent itemset is also infrequent. Frequent itemsets mining is a core component of data mining and variations of association analysis, like association-rule mining and sequential-pattern mining. In frequent itemsets are produced from very big or huge data sets by applying some rules or association rule mining algorithms like Partition method, Apriori technique, Incremental, Border algorithm Pincer-Search, and numerous other techniques that take larger computing time to compute all the frequent itemsets. Extraction of frequent itemsets is a core step in many association analysis techniques. An itemset is known as frequent if it presents in a large-enough portion of the dataset. This frequent occurrence of item is expressed in terms of the support count. Therefore, it needs complicated techniques for hiding or reforming users' private information during a data gathering process.Data Mining is the process of finding correlation or patterns among dozens of fields in large relational databases. It is to extract interesting information or patterns from data in large databases. Data mining is the procedure for discovering data from different viewpoints and summarizing it into valuable information. This information can be used to improve costs and profits of data information or both. Data mining is processed with the great deal of consideration in the information construction and in society recently, because of the extensive preventability of huge amounts of data and the future necessitate for figuring such data into practical information and acquaintance.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

II. LITERATURE SURVEY

ShipraKhare et al [1] this system proposed a method forin data mining and knowledge discovery technique areas, frequent pattern mining plays an important role but it does not consider different weight value of the items. The frequent itemsets are patterns or items like itemsets, substructures, or subsequences that come out in a data set frequently or rapidly. The frequent itemsets are patterns or items like itemsets, substructures, or subsequences that come out in a data set frequently or rapidly. In this paper presenting review of various frequent pattern mining and weighted itemset mining.

SujathaKamepalli et al [2] describedInfrequent Weighted Association Mining (IWAM) is one of the main areas in data mining for extracting the rare items in high dimensional datasets. Traditional Association rule mining algorithms produce large number of candidate sets along with the database scans. Due to large number of transactions and database size, traditional methods consume more time to find the relevant association rules with the specified threshold. Prior and post database scans are required an additional effort to validate the association rules. Most of the existing weighted models are implemented for mining frequent itemsets, but finding infrequent itemset mining are useful in many recent fields like web, medical, cloud, complex databases, protein sequence etc. In weighted infrequent association rule mining, each item in the transaction is assigned a weight in order to mine high utility infrequent itemsets. In this proposed work, weighted association rule mining algorithm is proposed to find infrequent itemsets using weighted threshold measures. Proposed approach gives better results on real-time datasets compare to existing weighted models.

J.Jaya et al [3] have approaches Itemset mining is a data mining method extensively used for learning important correlations among data. Initially item sets mining was made on discovering frequent item sets. Frequent weighted item set characterizes data in which items may weight differently through frequent correlations in data's. But, in some situations, for instance certain cost functions need to be minimized for determining rare data correlations. Determining these types of data is more challenge and interesting research than mining frequent data in items. This paper surveys various methods for frequent itemset and infrequent item set mining of data. This work differentiates various methods with each other during mining of data. Finally, comparative measures of each method are presented which provides the significance and limitations of frequent and infrequent mining of data in item sets.

III. METHODOLOGY

The system design considerations divide the system into many subsystems depending on the requirements of the system. The system design tells about the overall system architecture and is concerned with identifying various system components. System design states relationships between components and also the software structure. It maintains a design decisions record and provides a plan for the implementation phase.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

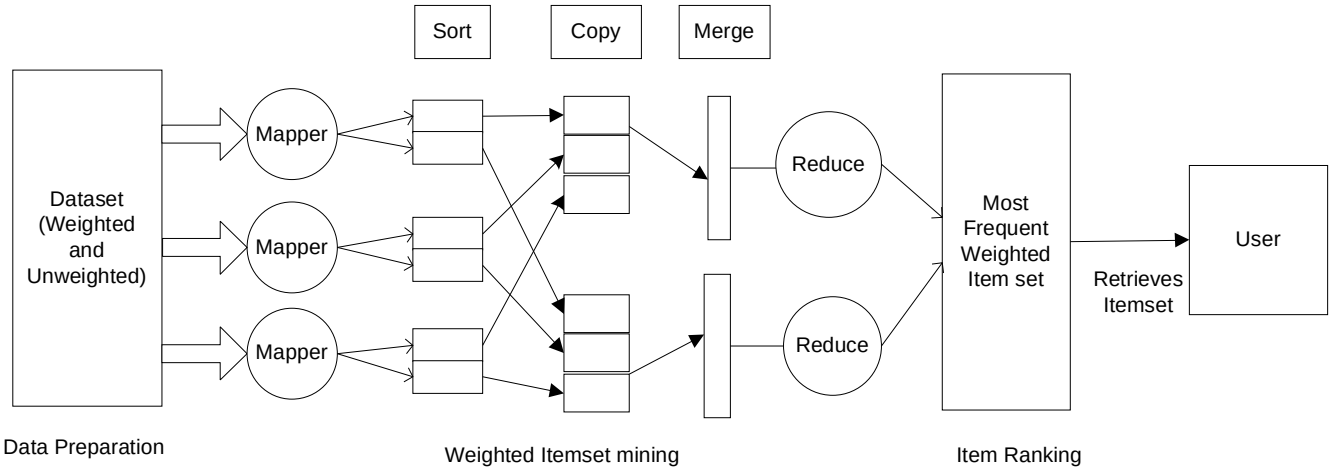


Figure 1: System architecture

a. Mapper function

The mapper function forms the input data as document or registry. Input data is put away in the Hadoop framework. The info record is then sent to the mapping process. The map function partitions the entire information into sub data. Every guide will prepare the subdata.

Algorithm for Map ()

Input: Split-Si;
Output: <key1, value1>;
key: frequent weighted k-Itemset of the split Si;
Value: local FWI-support.
Begin
For each weighted transaction Tw in Si.
For each weighted ItemsetIw in tw. /* Iw is all possible subsets of tw.*/
Wsupp = Cal_FWI-support (Iw) /* Calculate IWI-Support of Iw in the transaction (Tw) */
Output
(Iw,wsupp)
End for
End for
End;

Algorithm for Reduce ()

Input: <key1, value1>;
key1: local weighted candidate Itemsetof the split Si;
value1: local FWI-Support, weighted minimumsupport (wms).
Output: <key2, value2>;
key2: frequent weighted Itemsets;
value2: global IWI-Support.
Begin
ForeachIw in key1 do
val2= Cal_global FWI-support(Iw) /* Calculate IWISupport
ofIw in the entire dataset */
If(val2>wms)then

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

output(Iw , val2)
End if
End for
End

b. *ECC Algorithm*

Elliptic bend cryptography (ECC) is algorithm which deals with public key cryptography in light of the algebraic structure of elliptic curve over limited fields.Elliptic curves are appropriate for encryption, advanced marks, pseudo-arbitrary generators and different assignments. They are utilized as a part of integer factorization algorithm that has applications in cryptography, for example, Lenstra elliptic bend factorization. In this project we are using Elliptic Curve Integrated Encryption Scheme (ECIES), which is also called as Elliptic Curve Encryption Scheme used for encryption to give semantic security against an adversary who is permitted to utilize chosen plaintext and picked cipher text attacks.

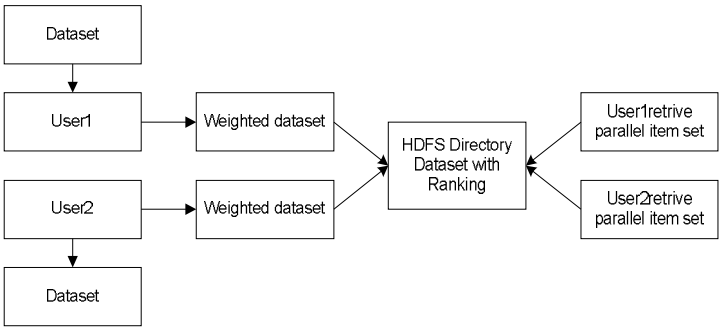


Figure 2: Dataflow diagram

2.1 Parallel Weighted Itemset Mining from BigData

Parallel Weighted Itemset Miner (PaWI) is a new datamining environment aimed to analyze Big Data equippedwith item weights. The main environment blocks are brieflyintroduced below. A more detailed description is given inthe following sections.

1 Data preparation

This step entails preparing data to the subsequent itemsetmining process. The source data is acquired, stored in atransactional dataset, and equipped with item weights. A*transactional dataset* is a set of transactions. Each*transaction* is a set of (not repeated) *items*.A transactional dataset whose items are enriched with weights are known as *weighted transactional dataset*.A weighted item is a pair (*item* and *weight*) . Depending on thecontext of analysis, items may represent different concepts(e.g., products, objects, places, stocks).For example, let us consider the dataset reported inTable I. It is an example of weighted transactionaldataset consisting of five transactions, each one representinga different customer of a e-commerce company. For eachcustomer the list of purchased items is known. For instance,customer with id 1 bought items *A with weight 2*, *B with weight 4*, and *C with weight 1*. Note thateach transaction, which represents a distinct electronic basket, may contain an arbitrary number of items.

Example of Weighted dataset: Item Bought by customers

Customer Id	Purchased Items
1	A2,B4,C1
2	A3,B5,E1
3	A1,B2,C3
4	A5,B1,E2
5	A2,D2,C1

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

2 Weighted itemset mining

This step focuses on mining frequent weighted itemsetsfrom the prepared weighted dataset. A *k*-itemset(i.e., an itemset of length *k*) is a set of *k* items. Thetraditional support value of an itemset in a transactionaldataset is given by its frequency of occurrence in the sourcedataset.For example, *{A,B}* is an itemset indicating the cooccurrenceof items *A*and *B*. If we disregard item weights,this itemset has a support equal to 4 in Table I because itoccurs in four out of five transactions, meaning that most ofthe users purchased items *A*and Btogether.

Definition 1: Weighted support. Let *D* be a weightedtransactional dataset, *I* be a weighted itemset,(i_j, w_j)be aarbitrary weighted item such that $i_j \in I$. Let *T* (*I*) be thesubset of *D*'s transactions containing all the items in *I* and*f* an arbitrary aggregation function defined on item weights.The weighted support of *I* in *D* is defined as

$$wsup(I,D) = \sum_{t_q \in T(1)} f_{j,kz|}\, i_{j,i_{k,\ldots,i_{z \in I}}}(w_j, w_{k,\ldots,} w_z)$$

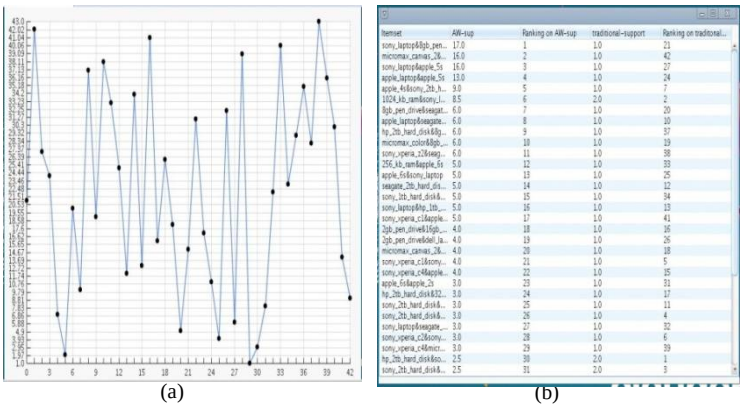
The weighted support is the summation of all the itemsetaggregation weights derived by the aggregation function *f*for every transaction in *T*. An arbitrary aggregation function*f* (e.g., min, max, average, and mode) can be potentially appliedto aggregate item weights within each transaction. Thechoice of *f* depends on the considered use cases. Hereafter, we will consider *f*=min (i.e., the least weightof any item in *I* is considered), because, as discussed inSection IV, the selected patterns are deemed as particularly useful for analyzing real Big datasets.

3 Item-set ranking

The manual exploration of all the itemsets (weighted ornot) mined from Big data is practically unfeasible. Hence, tosupport the knowledge discovery process experts may wouldlike to access only a subset of most interesting patterns. Thisstep focuses on ranking the mined itemsets according to theirlevel of significance in the analyzed data.To filter and rank the mined itemsets, the support measureis the most commonly used quality index. To copewith weighted data, for each candidate itemset the PaWIsystem computes both the traditional and weighted supportmeasures.

IV. **RESULTS**

After compilation of mining part the graph will be generated according to the result produced by the mining. The below fig shows the generation of graph, and is the comparison between the average weight support and tradition weight support by assigning the rank according to the weight. AW-sup is arranged in the ascending order whereas tradition support is arranged randomly.



(a) (b)

Figure 3: represents (a)Generating of graph; (b)Ranking based on AW_Support and Traditional_Support

V. **CONCLUSION**

This paper concludes a disseminated and parallel result for the issue of obtaining frequent items from huge ranked datasets. The suggested framework, executing on a Hadoop framework, conquers the constraints of best in class

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

approaches in adapting to datasets enhance with item weights. The experiment, executed on a dataset, confirms the significance of the mining result in real content.

REFERENCES

[1] ShipraKhare , Prof. Vivek Jain, "A Review on Infrequent Weighted Itemset Mining Using Frequent Pattern Growth", Vol. 5 , 2014
[2] SujathaKamepalli, Raja SekharaRaoKurra, SundaraKrishna.Y.K, "Infrequent Weighted Item Set Mining in Complex Data Analysis", Volume 103 – No.5, October 2014.
[3] Kalaiyarasi. P1, Manikandan. M, "Clustering Based Infrequent Weighted Itemset Mining", Volume No.03, Special Issue No. 02, February 2015.
[4] SeemaVaidya and Deshmukh PK, "Predicting Rare Disease of Patient by Using Infrequent Weighted Itemset'.
[5] Sakthi Nathiarasan1, Kalaiyarasi2, Manikandan3, "Literature Review on Infrequent Itemset Mining Algorithms", International Journal of Advanced Research in Computer and Communication EngineeringVol. 3, Issue 8, August 2014
[6] T Ramakrishnudu, "Mining Interesting Infrequent Itemsets from Very Large Data based on MapReduce Framework".
[7] Haifeng Li, Ning Zhang, Zhixin Chen, "A Simple but Effective Maximal Frequent Itemset Mining Algorithm over Streams",Vol. 7, No. 1, January 2012.
[8] Shabnam, Shashikala M K, "A Scalable Approach to FIM by Means of MR" Volume-5, Issue-5.
[9] R. Lakshmi Prasanna , Dr. G.V.S.N.R.V. Prasad, "Infrequent Weighted Item Set Mining Using Frequent Pattern Growth", International Journal of Engineering & Science Research, Vol-5/Issue-11
[10]KaramjitKaur, Rajeev Bedi, R.C.Gangwar, " Comparative Analysis Of Non-Frequent Pattern Mining Approach", International Journal Of Technology Enhancements And Emerging Engineering Research, Vol 3, Issue 06 65.
[11]Ying Liu, Wei-keng Liao,AlokChoudhary, "A Fast High Utility Itemsets Mining Algorithm".