Map Reduce: Fast Parallel Association Rule for Big Data

Monika N, Mr. Prakash V Parande
M.Tech, Dept Of CSE Visvesvaraya Institute of Advance Technology, Bengaluru, India.
Assistant Professor, Dept Of MCA, Visvesvaraya Institute Of Advance Technology. Bengaluru, India.

Abstract— The data mining technique in this work that is subjugated by taking out the repeated re-occurrences among data items is frequent item set mining. The items in the many applications are enhanced with weights which denote the significance in the explored data and push the weights of items to the mining process. The distinct feature in this work includes mining of frequent item sets, that has been an essential part of data analysis and data mining. The method specifically reports the difficulty of the item sets that are provided with respective weights from data augmented source which has pre-assigned weights. Disparate this has focus on the planning a parallel and distributed method that is capable to handle with the major weighted datasets. The proposed system provides a mining method with items that is provided with weights that is centered on the map reduce standard. To determine its action and scalability, the suggested method was verified with actual larger datasets. This project profits e-commerce customers, where companies may consider as interesting to effort preferment and publicizing, market basket study, census data exploration, and text summarization

Keywords—Data mining, Frequent item set mining

INTRODUCTION

The current architecture deals with large amount of data due to the fast advancement of the industrial technologies which leads to extensive, assorted and rapid data storage that is difficult to handle. This includes planning and promotions of commercial companies necessitate stock up and know the items that are purchased by the customer. To support this there exist a need for the big data gathering Data Mining solutions including clustering algorithms and classification algorithms that gauge towards big data. The frequent data mining is a data mining technique which is used to determine correlation among data items. It include mining algorithm that push items into the mining process these items are based on the significance weight each item set and considered as particular local significance that is associated with it. This technique had found applications in market basket analysis, census data analysis. This work includes large scale-scale item set mining that scale towards the large datasets and include parallel and distributed item set mining. The weighted item set mining includes association rules that are enriched with weights indicating the item significances. Each item in the transaction database contains many components and most of the association rules do not consider few components such as quantity.

A. Statement of problem

The data that is collected from the large transaction dataset contains the review of items that are purchased by the customer and includes duplicate entry of same items that are purchased this leads to the slowdown of the entry of items that are sold out at high rate which leads to reduce in the ordering of stock. This can be easily achieved by providing weights to the items and map reduce approach is used to remove duplicates and filter the dataset. AW-SUP values are calculated based on the item set ranking. The values with highest AW-SUP will be ranked first. This work addresses to the extracting frequent item set from the large datasets. Many of the in-memory algorithms do not provide equivalent and dispersed solution for above given process, to overcome this parallel weighted miner for each item set is introduced which relies on parallel and distributed implementation running on hadoop. The above process is made scalable by Map Reduce model that is introduced. This model can process large datasets this also provides many map functions producing many item sets with key-value pairs.and reduce function combines the values associated with the same intermediary key.

B. Objective

To collect the weighted data set from the provided source such as Amazon, Flip kart etc, and to provide weight to each items. To store the items into the HDFS directory for mining purpose. To prepare the data for removing duplicate rows in the weighted item set and to filter the data. To calculate the average weighted support value by partial mining and based on this the items are ranked the items with highest AW-SUP values will be given the first priority.

This work propose Parallel Weighted Item set miner, a new parallel and distributed framework to extract frequent weighted item sets from potentially very large transactional datasets enriched with item weights.

The framework relies on a parallel and distributed-based implementation running on an Hadoop cluster. To make the mining process scalable towards Big Data, most analytical steps performed by the system are mapped to the Map Reduce programming paradigm.

II. BACKGROUND

The mining process includes a large-scale item set mining and weighted item set mining which include Frequent item set and association rule mining are broadly trial data mining methods which were first introduced in the large-scale item set mining. Many of the parallel and distributed item set mining approaches are used to move towards larger datasets, for this purpose an Apriori-based mining is used. The Aprori-based approach provides less scalability when compared to the other algorithms such as projection-based on the larger datasets. In the traditional item set mining tasks the items fitting to each and every transaction of the datasets are considered equally which then addresses to the association rules this is the technique used to determine the customer purchasing details from the transaction datasets that fulfills assurance and self-confidence.

III. LITERATURE SURVEY

The data that is accessed from the search engines used to process some tera bytes of data this data is complex to cope up with because of high speed and heterogeneous in nature developed by D. Agrawal, S.Das and A.EI Abbadi [1] the Database management systems that comprise update intensive application workloads and decision support systems that are used for the explanatory and deep analytics are a significant part of the cloud infrastructure provides many transactions from traditional systems to the next generation cloud. This includes the systems for supporting larger applications and for ad-hoc analytics and decision support, since the database management comes up with many challenges as above. This work includes system analysis in detail and also to fall into place of designing the choices in the larger database systems.

The data that is copied by the analysts has the need for the data mining solutions in studying and analyzing the weighted and un weighted item set mining process developed by H. T. Lin and V. Honavar [7] the appearance of many data that is linked together physically distributed and in parallel maintained RDF stores provides unparalleled opportunity for predictive modeling and knowledge discovery from the above provided data. Nevertheless the existing machine learning methods are inadequate in their applicability because it is neither attractive nor feasible to bring the data in a centralized location analyze in terms of the access, memory, bandwidth, computational restrictions, privacy and confidentiality constraints.

The RDF data store brings in arithmetical query based formulations of several representative algorithms. It introduces a dispersed learning framework that form a chain by the interlinked data store. This work also includes novel applications of metric reconstruction technique and also provides data fragmentations.

The data mining technique used in this work is the frequent item set mining which has applications in the customer transactions developed by R. Agrawal, T. Imielinski, and Swami [5] this work includes the different transactions made by the customers in the larger transaction databases where each transaction consists of items purchased by a customer in a visit. It also presents an efficient algorithm that generates all significant association rules between items in the database that incorporates buffer management and novel estimation and pruning techniques where the above algorithm is applied to the large retailing company having the sales data that gives effectiveness to the algorithm.

.

IV. PROPOSED SOLUTION

The Map Reduce standard is used where each and every item in the transaction database is provided with respective weights that indicates the particular ratings to every item that are purchased by the customer.

The source data is prepared by acquiring the data that is enriched with the weight is then transferred to the mining process using technique called as data filtering and the result of this is stored in the HDFS data repository. The weighted item set mining the frequent weighted and un weighted item sets are extracted from the prepared datasets this is based on parallel and distributed item set mining which run on the hadoop cluster.

The item set ranking process the outcome of the weighted and un weighted item set mining and these are compared with each other and best pattern is selected based on a new quality measure which combines traditional with weighted support counts.

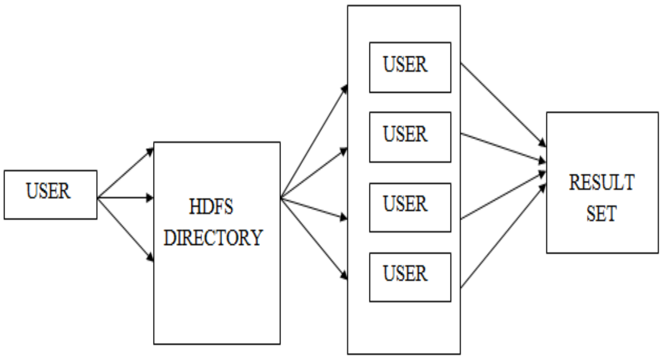


Fig.1. Weighted item set mining architecture

FIG 1 shows the overview of weighted item set mining architecture where the user uploads the review to the HDFS directory which consist of blocks that performs file system maintenance. The HDFS directory consists of weighted and un weighted datasets, weighted item sets indicates average level of interest of items within each transaction. The items will be ranked by comparing weighted and un weighted results and then it is parallel retrieved by the user then it is mapped to the result set.

A. *Sign in module*

The user sign in to the larger dataset in the mining process with the username and the password only after successful login he or she perform the required operations.

B Sign-up module

The user has to register his or her details in the cloud database with the name, password, email-id etc before his or her login.

C. User module

After the successful login the user can insert, generate and upload the data where the source data is prepared by acquiring the data that is enriched with the weight is then transferred to the mining process.

D. Cloud module

This module has the data server where user can perform the basic operation and weighted and un weighted item sets are extracted from the prepared datasets and best one is ranked..

E. Insert data module

In this module the required data will be selected by the user and these selected items will be inserted into the HDFS directory for the mining process.

F. Big Fim algorithm

- Step1:
  First collect the data sets from transaction database.
- Step 2:
  Select the required data from datasets.
- Step3:
  Store the dataset with the provided weights.
- Step4:
  The mapping part will be introduced in mining.
- Step5:
  Search the parallel items in datasets.
- Step6:
  In reducing part we can change the result.

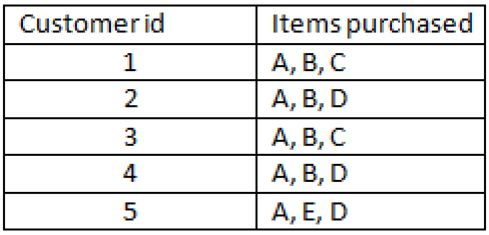
The Big Fim algorithm works as follows here the users collect the data sets from transaction data sets and then the user uploads the review in to the HDFS directory this consists of weighted and un weighted item sets, the weighted item sets will be ranked by comparing the results of item sets. The mapping part will search the parallel items in the data sets then it will be mapped to the result set.

V. *Parallel Weighted item set mining from larger datasets*

The Parallel Weighted item set miner is the technique used to study item weights here the data with item weights that are acquired and transformed to mining process using data filtering technique this result will be stored in HDFS directory. In the weighted item set mining process all the weighted and un weighted items are extracted from prepared datasets. These items are ranked by comparing the weighted and un weighted item sets with each other and the most interesting patterns are selected based on the new quality measure.

A. Data prepation

This step entails preparing data to the subsequent item set mining process the data is acquired, stored in the transactional dataset, and provided with item weights. consider dataset that is given in Table 1, that consists of transaction that are made by the customers. Here A,B, C, D, E denotes the items that are purchased by the customer



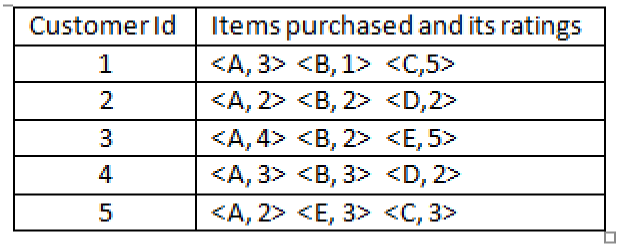
Customer id	Items purchased
1	A, B, C
2	A, B, D
3	A, B, C
4	A, B, D
5	A, E, D

Table I. Un weighted datasets

B. Weighted item set mining

This step shows mining frequent weighted item sets from the prepared weighted dataset. . A weighted item is a pair of < item, weight > where each item is associated with weight Table II shows customer with id 1 rated item *A* as 3, item *B* as 1, and item *C* as 5. The traditional support value of an item set in a transactional dataset is given by its frequency of occurrence in the source dataset the *{A, B }* is an item set indicating the co occurrence of items *A* and *B* . If we disregard item weights, this item set has a support equal to 4 in Table I because it occurs in four out of five transactions, this shows that most of users purchased items *A* and *B* together. If we disregard item weights, this item set has a support equal to 4 in Table I because it occurs in four out of five transactions, mean that most of the users purchased items *A* and *B* together.

TABLE II. *weighted datasets*



Customer Id	Items purchased and its ratings
1	<A, 3> <B, 1> <C,5>
2	<A, 2> <B, 2> <D,2>
3	<A, 4> <B, 2> <E, 5>
4	<A, 3> <B, 3> <D, 2>
5	<A, 2> <E, 3> <C, 3>

C. Item set ranking

The item sets mined from Big data is practically unfeasible. This step focuses on ranking the mined item sets according to their level of significance in the analyzed data. To filter and rank the mined item sets, the support measure is the most commonly used quality index . To cope with weighted data, for each candidate item set the weighted item set system computes both the traditional and weighted support measures. While the traditional support value indicates the observed frequency of occurrence of the considered combination of items in the source dataset, in weighted support counting item set occurrences are weighted by the least item weight . The item sets are selected by combining the weighted and traditional support measure in a new measure called average weighted support.

VI. RELATED WORKS

Parallel Database management system aims at this debate by presenting an analysis of the advantages and disadvantages of the different approaches. The deep analytics include the new class of data applications such as complex statistical analysis and machine learning techniques on huge amount data to garner intelligence from data. Scalable data management focus on class of systems that are designed to update heavy web-applications deployed in the cloud.

The emergence of many interlinked, physically distributed, and autonomously maintained RDF stores such as the cloud offers unprecedented opportunities for predictive modeling and knowledge discovery from such data. However existing machine learning approaches are limited in their applicability because it is neither desirable nor feasible to gather all of the data in a centralized location for analysis due to access, memory, bandwidth, computational restrictions, and sometimes privacy or confidentiality constraints.

VII. CONCLUSION

The above proposed schema provides distributed and parallel approach for the crisis of retrieving items that are common in dataset and these items will be provide with particular ranks based on these ranks the items will be prioritized, these ranks are used to know which item will be sold quickly from the stock so that items can be ordered. The items with higher average support will be given the first priority. The mining association rule that is running on the hadoop cluster handles mining in the large datasets, the parallel mining provides the heterogeneity on the hadoop cluster provides progressive in the hardware by increasing the efficiency of the cache and also increases the system efficiency by data placement in hadoop clusters.

ACKNOWLEDGMENT

I thanks to my guide Mr. Prakash V Parande, Assistant Professor, Dept. of MCA, VTU-CPGS, Bengaluru Region, VIAT, Muddenahalli, for his valuable guidance throughout this research work.

REFERENCES

1) A. El Abbadi, “Cloud computing and Big data: current state and future opportunities,” on database technology New York, NY, USA: ACM, 2011.

2) M. Govindaraju, J. Hartog “Configuring a map reduce framework for performance heterogeneous clusters “ in 2014 IEEE International Congress on Big data, Anchorage, AK, USA, June 27 – July 2014.

3) S. Ghemawat and J. Dean “Mapreduce: simplified data processing on clusters,” in 16th conference on operating systems in 2012.

4) T. Imielinski, R. Agarwal, “Mining association rules between itemsets in large datasets,” in ACMSIGMOID 1993.

.

5) V.Honavar, “Learning classifiers from chain of multiple interlinked RDF data stores,” in IEEE International congress on Big data, Big data congress 2016.

6) L. Cagliero and P. Garza, “Item set generalization with cardinality-based constraints,” *Information Sciences*, vol. 244.

7) R. Agrawal and R. Srikant, “Fast algorithms for mining association rules in large databases,” in *VLDB’94, of 20th International Conference on Very Large Data Bases*, J. B. Bocca, M. Jarke, and C. Zaniolo, Eds. Morgan Kaufmann, 1994.

8) J. Han, J. Pei, and Y. Yin, “Mining frequent patterns candidate generation,” in *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, 2000.

9) [9] M. J. Zaki, S. Parthasarathy, M. Ogihara, and W. Li, “New algorithms for fast discovery of association rules,” in *International Conference on Knowledge Discovery and Data Mining (KDD)*, Aug 1997.

10) S. Moens, E. Aksehirli, and B. Goethals, “Frequent item mining for big data,” in *SML: Big Data 2013 Workshop on Scalable Machine Learning*. IEEE, 2013.

11) H. Li, Y. Wang, D. Zhang, M. Zhang, and E. Y. Chang, "parallel fp-growth for query recommendation,” *of the 2008ACM conference on Recommender systems*, serRecSys ’08. New York, NY, USA: ACM, 2008.

12) W. Wang, J. Yang, and P. S. Yu, “Efficient mining of weighted association rules (WAR),” in *Proceedings of the SIGKDD international conference on Knowledge discover and data mining, KDD’00*, 2000, pp. 270–274.

13) F. Tao, F. Murtagh, and M. Farid, “Weighted association mining using weighted support and significance framework, "In *Proceedings of the ninth ACM SIGKDD international on Knowledge discovery and data mining, KDD’03*, 2003, pp. 661–666.

14) K. Sun and F. Bai, “Mining weighted association rules preassigned weights,” *IEEE Transactions on Knowledge Eng.*, vol. 26, no. 4, pp. 903–915, 2014.[Online]Available:http://doi.ieeecomputersociety.org/10.1109/TKDE.2013.69 *and Data Engineering*, vol. 20, no. 4, pp. 489 –495, 2008.

15) P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction data Mining*. Addison Wesley, 2005.*conference on Symposium on Opearting Systems Design Implementation - Volume 6*, ser. OSDI’04, 2004.

16) [18] M. Zaki, “Parallel and distributed association mining: a survey,” *Concurrency, IEEE*, vol. 7, no. 4, pp. 14–25, Oct 1999.