A Survey on an effective incremental mining algorithm for frequent item-sets with big data based on MapReduce framework

[1] Shreerajsinh R Jadeja, [2] Prof. Satvik Khara, [3] Prof. Rikin Thakkar
[1] Student, [2] HOD, [3] Assistant Professor
Computer Engineering Department
Silver Oak College of Engineering & Technology, GTU, Ahmedabad, India
[1]shreeraj.rz@gmail.com, [2]svkhara@gmail.com, [3]rikinthakkar1991@gmail.com

***Abstract*— Due to the increasing use of very large databases and data warehouses, mining useful information and helpful knowledge from transactions is evolving into an important research area. Thus, most of the classic algorithms proposed focused on batch mining, and did not utilize previously mined information in incrementally growing databases. In the past, researchers usually assumed databases were static to simplify data mining problems. This research presents a new scalable algorithm Delta+ for discovering closed frequent item sets. Exploits a divide-and-conquer approach. Adopts several optimizations aimed to save both space and time in computing item set closures and their supports. Propose a new effective and memory-efficient pruning technique. Algorithm is scalable and outperforms algorithms like FP-Growth, in some cases even better. The performance improvements become more and more significant as the support threshold is decreased.**

***Keywords* – Big Data; Incremental Mining; Frequent item-sets; Incremental data mining algorithm; Map-Reduce framework;**

INTRODUCTION

Data mining is the practice of automatically searching large stores of data to discover patterns and trends that go beyond simple analysis. Data mining uses sophisticated mathematical algorithms to segment the data and evaluate the probability of future events. Data mining is also known as Knowledge Discovery in Data (KDD). The key properties of data mining are: Automatic discovery of patterns, Prediction of likely outcomes, Creation of actionable information, Focus on large data sets and databases.

There is a great deal of overlap between data mining and statistics. In fact most of the techniques used in data mining can be placed in a statistical framework. However, data mining techniques are not the same as traditional statistical techniques.
Traditional statistical methods, in general, require a great deal of user interaction in order to validate the correctness of a model. As a result, statistical methods can be difficult to automate. Moreover, statistical methods typically do not scale well to very large data sets. Statistical methods rely on testing hypotheses or finding correlations based on smaller, representative samples of a larger population. Data mining methods are suitable for large data sets and can be more readily automated. In fact, data mining algorithms often require large data sets for the creation of quality models.

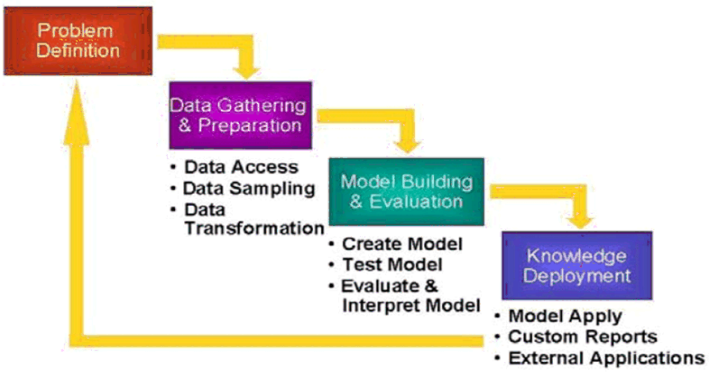

Figure.1

Association is a data mining function that discovers the probability of the co-occurrence of items in a collection. The relationships between co-occurring items are expressed as association rules.

Confidence

The confidence of a rule indicates the probability of both the antecedent and the consequent appearing in the same transaction. Confidence is the conditional probability of the consequent given the antecedent. For example, cereal might appear in 50 transactions; 40 of the 50 might also include milk. The rule confidence would be:

Association rules are often used to analyze sales transactions. For example, it might be noted that customers who buy cereal at the grocery store often buy milk at the same time. In fact, association analysis might find that 85% of the checkout sessions that include cereal also include milk. This relationship could be formulated as the following rule. Cereal implies milk with 85% confidence. This application of association modeling is called **market-basket analysis**. It is valuable for direct marketing, sales promotions, and for discovering business trends. Market-basket analysis can also be used effectively for store layout, catalog design, and cross-sell. Association modeling has important applications in other domains as well. For example, in e-commerce applications, association rules may be used for Web page personalization. An association model might find that a user who visits pages A and B is 70% likely to also visit page C in the same session. A and B imply C with 70% confidence.

Incremental mining

Data mining association rules are often done by computing the association rules for the whole source database. An incremental data load is a method of updating the dataset in which only new or modified records are uploaded to the project. If you perform a full data load, any records that were in the dataset and are absent in the new data no longer appear in the updated dataset.

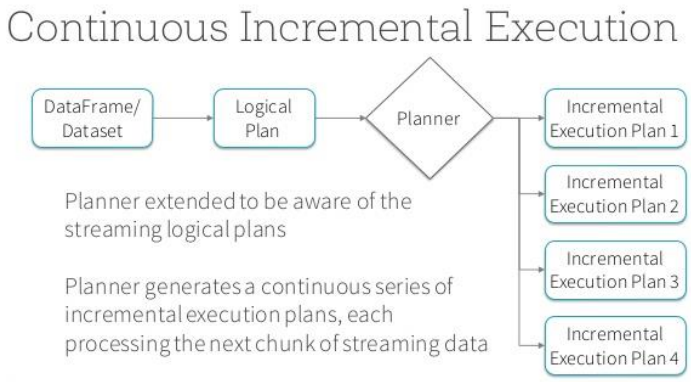

Figure.2

Hadoop Map-Reduce Framework

Hadoop Map-Reduce is a software framework for easily writing applications which process vast amounts of data (multi-terabyte data-sets) in-parallel on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner. Map-Reduce program executes in three stages, namely map stage, shuffle stage, and reduce stage.

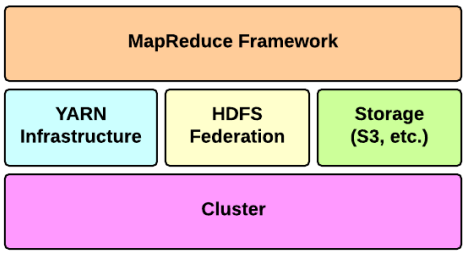

Figure.3 [22]

Map stage: The map or mapper's job is to process the input data. Generally the input data is in the form of file or directory and is stored in the Hadoop file system (HDFS). The input file is passed to the mapper function line by line. The mapper processes the data and creates several small chunks of data [22]

Reduce stage: This stage is the combination of the Shufflestage and the Reduce stage. The Reducer's job is to process the data that comes from the mapper. After processing, it produces a new set of output, which will be stored in the HDFS [22]

Combiner stage**:** When a MapReduce Job is run on a large dataset, Hadoop Mapper generates large chunks of intermediate data that is passed on to Hadoop Reducer for further processing, which leads to massive network congestion [22]

BACKGROUND THEORY

Incremental data mining is very important to solve the temporal dynamic property of knowledge, improve the performance of mining processes and efficiency of mining results. Incremental data occurs with the passage of time. Evolutionary methods can be adopted to solve such increment. A multiply evolutionary model is built to describe incremental data evolutionary mining processes. Copy operator, cross operator and mutation operator are designed. A general algorithm for dynamic evolutionary mining is also presented. The evolutionary incremental data mining method could solve the scalable problem of data mining better, and have high accuracy and good time performance.

In fact, the processes for dynamic data, especially incremental data, were considered in early data mining research. The literature [3] and [4] presented an iterative and interactive algorithms, which the sample data was selected from the whole data set and when new data added to the data set and the sample data selected again. If the data set is very large the iterative processes need consume much time and space resources. Based on Gold's work [6] the literature [5] presented the learning theory and technology of mining incrementally concepts through iterative learning, bounded instance space reasoning and k-feedback identity. The literature [5] mainly deeply studied the theory of incremental concept learning, and there were no experiment analysis from the points of practical use. With the deeply research and the wide applications, the researchers and practitioners gradually attach importance to the dynamic data and more and more algorithms involved in incremental data. The literature presented an incremental association rules discovery system. The system designed a meta-data frames model Agent and an objective frames model Agent. These frames models consisted of static classes and active classes. The method in literature can effectively discover association rules of incremental data. The literature [8] presented a Hybrid Distribution algorithm which can parallel deal with incremental data and scan data set only once. So far most the methods of mining dynamic data mainly concentrate on association rules discovery. Other algorithms about this mainly include web access patterns mining[9], data set partition methods[10], etc.

High Profit and High Utility Pattern Mining

Frequent item-set mining has some important limitations**.** The problem of frequent item-set mining is popular. But it has some important limitations when it comes to analysing customer transactions. An important limitation is that purchase quantities are not taken into account. Thus, an item may only appear once or zero time in a transaction[15]. Thus, if a customer has bought five breads, ten breads or twenty breads, it is viewed as the same. A second important limitation is that all items are viewed as having the same importance, utility of weight. For example, if a customer buys a very expensive bottle of wine or a cheap piece of bread, it is viewed as being equally important. Thus, frequent pattern mining may find many frequent patterns that are not interesting. For example, one may find that {bread, milk} is a frequent pattern. However, from a business perspective, this pattern may be uninteresting because it does not generate much profit. Moreover, frequent pattern mining algorithms may miss the rare patterns that generate a high profit such as perhaps {caviar, wine}

High Utility Itemset Mining

Unit Profit	
A	2
B	3
C	1
D	3
E	4
F	4
G	8

Transactional Database		Total
T_1	A(4), B(2), C(8), D(2)	28
T_2	A(4), B(2), C(8)	22
T_3	C(4), D(2), E(2), F(2)	26
T_4	E(2), F(2), G(1)	24

$minimum\ utility\ threshold\ \theta = 30$

Is {A, C} a high utility itemset?

$$u(\{A,C\})\ in\ the\ transactional\ database$$
$$= \underbrace{(4\times 2+8\times 1)}_{T_1} + \underbrace{(4\times 2+8\times 1)}_{T_2} = 32\ > \theta \rightarrow HUI$$

Figure.4

High utility item-set mining

To address these limitations, the problem of frequent item set mining has been redefined as the problem of **high utility item-set mining**. In this problem, a transaction database contains transactions where purchase quantities are taken into account as well as the unit profit of each item. For example, consider the following transaction database.

transaction database with quantities

Trans.	items
T_0	a(1), b(5), c(1), d(3), (e,1)
T_1	b(4), c(3), d(3), e(1)
T_2	a(1), c(1), d(1)
T_3	a(2), c(6), e(2)
T_4	b(2), c(2), e(1)

unit profit table

item	unit profit
a	5 $
b	2 $
c	1 $
d	2 $
e	3 $

Figure.5 [21]

Consider transaction T3. It indicates that the corresponding customer has bought two units of item "a", six unit of item "c", and two units of item "e". Now look at the table on the right. This table indicates the unit profit of each item. For example, the unit profit of items "a", "b", "c", "d" and "e" are respectively 5$, 2$, 1$, 2$ and 3$. This means for example, that each unit of "a" that is sold generates a profit of 5$. The problem of high utility item-set mining is to find the item-sets (group of items) that generate a high profit in a database, when they are sold together. The user has to provide a value for a threshold called "minutil" (the minimum utility threshold). A high utility item-set mining algorithm outputs all the high utility item-sets, that is the item-sets that generates at least "minutil" profit. For example, consider that "minutil" is set to 25 $ by the user. The result of a high utility item-set mining algorithm would be the following.

High utility itemsets

{a,c} : 28$	{a,c,e}: 31 $
{a,b,c,d,e}: 25 $	{b,c} : 28 $
{b,c,d}: 34 $	{b,c,d,e}: 40 $
{b,c,e} : 37 $	{b,d} : 30 $
{b,d,e} : 36 $	{b,e} : 31 $
{c, e}: 27$	

Figure.6 [21]

For example, consider the item-set {b,d}. It is considered to be a high utility item-set, because it has a utility of 40$ (generates a profit of 40$), which is no less than the minutil threshold that has been set to 25$ by the user. Now, let's look into more detail about how the utility (profit) of an item-set is calculated. In general, the utility of an item-set in a transaction is the quantity of each item from the item-set multiplied by their unit profit. For example, consider the figure below. The profit of {a,e} in transaction T0 is 1 x 5 + 1 x 3 = 8 $. Similarly, the profit of {a,e} in transaction T3 is 2 x 5 + 2 x 3 = 16 $. Now, the utility of an item-set in the whole database is the sum of its utility in all transactions where it appears. Thus, for {a,e}, its utility is the sum of 8$ + 16 $ = 24$ because it appears only in transactions T0 and transaction T3.

Trans.	items
T_0	a(1), b(5), c(1), d(3), (e,1)
T_1	b(4), c(3), d(3), e(1)
T_2	a(1), c(1), d(1)
T_3	a(2), c(6), e(2)
T_4	b(2), c(2), e(1)

item	unit profit
a	5 $
b	2 $
c	1 $
d	2 $
e	3 $

Figure.7 [21]

Problem with high utility pattern mining

First, it may be more interesting from a practical perspective to discover item-sets that generate a high profit in customer transactions than those that are bought frequently.

Second, from a research perspective, the problem of high utility item-set mining is more challenging. In frequent item-set mining, there is a well-known property of the frequency (support) of item-sets that states that given an item-set, all its supersets must have a support that is lower or equal. This is often called the "Apriori property" or "anti-monotonicity" property and is very powerful to prune the search space because if an item-set is infrequent then we know that all its supersets are also infrequent and may be pruned. In high utility item-set mining there is no such property. Thus given an item-set, the utility of its supersets may be higher, lower or the same. For example, in the previous example, the utility of item-sets {a}, {a,e} and {a,b,c} are respectively 20 $, 24$ and 16$.

RELATED WORK

Incremental data mining is very important to solve the temporal dynamic property of knowledge, improve the performance of mining processes and efficiency of mining results. Incremental data occurs with the passage of time. Evolutionary methods can be adopted to solve such increment. A multiply evolutionary model is built to describe incremental data evolutionary mining processes. Copy operator, cross operator and mutation operator are designed. A general algorithm for dynamic evolutionary mining is also presented. The evolutionary incremental data mining method could solve the scalable problem of data mining better, and have high accuracy and good time performance.

Propose a novel incremental data mining algorithm based on FP-Growth, using the concept of heap tree to address the issue of incremental updating of frequent item sets. They have proposed a novel incremental data mining algorithm based on FP-Growth, using the concept of heap tree to address the issue of incremental updating of frequent itemsets. This method retains the advantages of FP-Growth, and significantly reduces the complexity associated with re-mining frequent itemsets during incremental updating.[1]

A novel Map-Reduce framework for an association rule algorithm based on Lift interestingness measurement (MRLAR) which can handle massive datasets with a large number of nodes. It discovers a set of co-location patterns using a GUI (Graphical User Interface) model in a less amount of time, as this application is implemented using a parallel approach-A Map-Reduce framework. [2]

The parallel scheduler is modeled for resource and task using particle swarm optimization to manage the assignments of map and reduce task. The Resource management is carried to manage a resource slot, which reduces the consumption of energy when running the application achieves optimal schedules. Performance evaluation of the frameworks is compared with state of approaches. The parallel scheduler is modeled for resource and task using particle swarm optimization to manage the assignments of map and reduce task. [3]

A Map-Reduce framework. This framework uses a grid based approach to find the neighboring paths using a Euclidean distance. The framework also uses a dynamic algorithm in finding the spatial objects and discovers co-location rules from them. Once co-location rules are identified, we give the input as a threshold value which is used to form clusters of similar behavior. [4]

To handle the massive amount of tweets we have used Hadoop Map Reduce framework to perform data mining analytic operations such as data cleansing, data classification and data clustering. Prediction model for the movie review is built by using Naïve Bayes Classifier and accuracy of the prediction is calculated with the help of binomial test as it conforms to the Bernoulli distribution. [5]

Growing size and complexity of data in data storage, distributed data mining algorithms has to be designed to handle Big Data in compatible with the latest technology available on distributed computing. Earlier research activities in data mining comprises, focus on increasing the performance for single task computing algorithms rather than distributed computing which would provide more fast and scalable environment for processing large datasets. [6]

CONCLUSION

We propose a new effective algorithm to analyze high profit item sets to achieve the rich benefit of market basket analysis. Since big data with map reduce framework is used, the possibilities are wide to get better results of large volume data which is having item sets that satisfy conditions for high utility item sets. The advantages over the existing algorithms will be verified. The proposed solution will overcome the limitations of the existing schedulers described here and gives more control to the users for Job execution.

ACKNOWLEDGEMENT

No task can be accomplished without proper guidance, support and appraisal. I am deeply thankful to many people who helped me either directly or indirectly for this work and provided their invaluable cooperation to me to complete it. At Last, I heartily offer my regards and blessings to all of those who supported me in any respect during the completion of my work.

REFERENCES

[1] A Novel Incremental Data Mining Algorithm based on FP-Growth for Big Data Hong-Yi Chang, Jia-Chi Lin, Shih-Chang Huang – IEEE-2016

[2] A Novel Mapreduce Lift Association Rule Mining Algorithm (MRLAR) for Big Data. Nour E. Oweis, Mohamed Mustafa fouad, Sami R oweis. IEEE-2016

[3] Efficient Hybrid framework for parallel Resource and Task scheduling in the Map reduce programming. S Hemalatha, S.Valarmathi – IEEE-2016

[4] A Map-Reduce Framework for finding clusters of Colocation patterns – A summary of results. M.Sheshikala, D. Rajeswara Rao, R. Vijaya Prakash – IEEE-2016

[5] An Efficient Framework of Data Mining and its Analytics on Massive Streams of Big Data Repositories, Disha D N, Sowmya B J, Chetan, Dr. Seema S. IEEE-2016

[6] Distributed FP-ARMH Algorithm in Hadoop Map Reduce Framework. Surendar Sountharrajan Sehar Assistant Professor (Senior Grade), IEEE-2013

[7] L. Hetland, "Listening to Music Enhances Spatial-Temporal Reasoning: Evidence for the" Mozart Effect"," Journal of Aesthetic Education, pp. 105-148, 2000.

[8] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," AI magazine, vol. 17, p. 37, 1996.

[9] E. Baralis, L. Cagliero, T. Cerquitelli, and P. Garza, "Generalized association rule mining with constraints," Information Sciences, vol. 194, pp. 68-84, 2012.

[10] R. Srikant and R. Agrawal, Mining generalized association rules: IBM Research Division, 1995.

[11] J. S. Park, M.-S. Chen, and P. S. Yu, "Using a hash-based method with transaction trimming for mining association rules," IEEE Transactions on Knowledge and Data Engineering, vol. 9, pp. 813-825, 1997.

[12] R. Agrawal, T. Imieliski, and A. Swami, "Mining association rules between sets of items in large databases," in ACM SIGMOD Record, 1993, pp. 207-216. [7] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in Proc. 20th int. conf. very large data bases, 1994, pp. 487-499.

[13] J. Han, J. Pei, Y. Yin, and R. Mao, "Mining frequent patterns without candidate generation: A frequent-pattern tree approach," Data mining and knowledge discovery, vol. 8, pp. 53-87, 2004.

[14] J. Guo and Y.-g. Ren, "Research on Improved A Priori Algorithm Based on Coding and MapReduce," in Web Information System and Application Conference (WISA), 2013, pp. 294-299.

[15] J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, et al., "Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth," in icccn, 2001, p. 0215.

[16] D. W. Cheung, J. Han, V. T. Ng, and C. Wong, "Maintenance of discovered association rules in large databases: An incremental updating technique," in Proceedings of the Twelfth International Conference on Data Engineering, 1996, pp. 106-114.

[17] Z. Zhou and C. Ezeife, "A low-scan incremental association rule maintenance method based on the apriori property," in Advances in Artificial Intelligence, ed: Springer, 2001, pp. 26-35.

[18] Oweis, N. E., Owais, S. S., George, W., Suliman, M. G., & Snášel, V. “A Survey on Big Data, Mining: (Tools, Techniques, Applications and Notable Uses)”. In Intelligent Data Analysis and Applications Springer International Publishing, pp. 109-119, 2015.

[19] Fouad, M. M., Oweis, N. E., Gaber, T., Ahmed, M., & Snasel, V. “Data Mining and Fusion Techniques for WSNs as a Source of the Big Data”. Procedia Computer Science, 65, ISSN 1877-0509, pp. 778-786, 2015.

[20] Ding, Guoru, Qihui Wu, Jinlong Wang, and Yu-Dong Yao. "Big Spectrum Data: The New Resource for Cognitive Wireless Networking.", arXiv preprint arXiv: 1404.6508, 2014.

[21] http://data-mining.philippe-fournier-viger.com/introduction-high-utility-itemset-mining/

[22] https://www.tutorialspoint.com/hadoop/hadoop_big_data_overview.htm