AN EFFICEINT APPROACH FOR MINING ASSOCIATION RULES FROM HETROGENOUS UNCERTAIN DATA STREAMS USING BIG DATA

1. Chinnapaga Ravi 2. M Bal Raju 3. N Subhash Chandra
1. Research scholar 2. Professor 3.professor
Computer science and engineering, JNTUH, Hyderabad, India

Abstract –Data Mining aims to search for implicit, previously known and potentially useful information from data. Big Data Mining is the capability of extracting useful information from large datasets or stream of data. The existing system attempts to search the pattern of interest from probabilistic database. However, the output sometimes includes the uncertain data from existential probabilities. In many real-life applications, users may look for a tiny portion of this large search space for Big Data Mining. The proposed system reduces the search space to a greater extent as it concentrates more on the constraints by using the Map Reduce model. The users are given complete freedom to express their interests by specifying their own constraints. Besides classification and clustering, anomaly detection, frequent pattern mining and association rule mining are included as the latter two analyze valuable data and helps the producer by finding the interesting or popular patterns that reveal customer purchase behavior. The algorithm proposed here greatly reduces the search space for Big Data mining of uncertain data, returning only those patterns that are interesting to the users for Big Data analytics.

Keywords—Big data models, Big data analytics, Frequent Patterns, Constraints, Uncertain Data

I. INTRODUCTION: Data mining mainly deals with extracting data from the data warehouse. The data warehouse contains very large amount of data. This large section will have basically 2 sets of data. They are the interested data and uninterested data. The uninterested data must not be displayed on the screen. The interested data is expected to be on the screen. There are different tools available to reduce the space to mine the data. This tool also increases the speed of the system. In other words, the whole of the result will be displayed in less amount of time and this will take less time and will be more efficient. As the day passes the more of the data gets collected. The "uncertain" data which is asked for must be searched in already existing data plus the newly added data. The uncertain data is the data given by the user. Since the system will have the least or no guess about the users next move, the data given will be uncertain to the system. In the traditional method whole of the data warehouse is searched for the data and hence the time required was too much. The data mining was not much in scene in the beginning of the computers. As the computers have evolved, the storage area is increased. Before the whole of the database itself dint contain so much of data as is in the present case's one database. The search space is drastically increased and is the need to increase the advancement in the search tools. Big Data Mining, in brief, is the intersection of big data and data analytics. The Big Data Mining is nothing but the capability of extracting useful information from large datasets or stream of data. If the data mining is done on very large set of data then it can be termed as "Big Data Mining" .The Big Data Mining is used in almost each and every field today. Without the use of data mining it will be very difficult to extract the required data. There are numerous search engines that are existing today. The entire search engine uses one or the other algorithm to extract the interested data. The field can be medical, banking, business, games, science and engineering and many more.

II. LITERATURE SURVEY

Database management tools are defined from 3Vs to 4Vs. 3Vs refers to volume, velocity, and variety. [1]Most of the presented approaches in data mining are unable to handle the large amount of data in a proper way. The key issue in the examination of huge information is the absence of coordination between database frameworks and in addition with investigation tools(for illustration information mining and factual examination).
The examination on many-sided quality hypothesis of enormous information will help the comprehension of fundamental attributes and development of complex examples in huge information, rearrange its portrayal, improves learning deliberation, and guide the plan of processing models and calculations on huge information. The difficulties of huge information examination are ordered into four principle categories. They are information stockpiling and investigation; learning revelation and computational complexities; versatility and representation of information; and data security. Information diminishment, Data determination, Feature choice is a fundamental errand particularly when managing expansive datasets. A standard process is to change the information that are semi structured and unstructured into organized and after that apply information mining calculations to remove learning. It is difficult to search the user interesting pattern among thousands of terabyte and hence allows users to express their interest in terms of constraints and uses the Map Reduce model to mine uncertain Big data for frequent patterns that satisfy the user-specified constraints.[2]As the technology advances ,the Big Data information explosion is mainly due to the vast amounts of data generated by social media platform ,data input from omni-channels ,various mobile devices ,user generated data ,multi-media data and so on. This lead into the new era of big data. In uncertain data each transaction contains items and their existential probabilities. Existing techniques are fp growth and Apriori algorithm. As implied by its name, Map Reduce involves two key functions: “map” and “reduce”. One of the problems to uncover hidden knowledge from Big Data is concept where statistical properties of the attributes and their target classes shift over time resulting in less accuracy.

III Implementation: -

Data Mining Data mining is the process of collecting and aggregating data from different perspectives and sources and analyzing the data to generate meaningful information. Association Rule Mining Association rule mining, introduced in 1993, is one of the most useful applications of data mining. Association rule mining makes it possible to discover patterns and interesting relationships between items in databases (Wakchaware, 2014). Classical Apriori Algorithm the Classical Apriori Algorithm (CAA), which is utilized for finding successive item sets, was produced by Aggrawal and Srikant in 1994. The CAA is extremely easy to actualize and comprises of two principle steps; the joint venture for producing hopeful item sets and the prune advance for dispensing with applicant item sets that are not visit (Kaur, 2014). The CAA is the traditional algorithm used to generate Association Rules. The CAA despite its simplicity and ease of implementation has several drawbacks. Some weaknesses of the algorithm include:

- the generation many of candidate itemsets consisting of many infrequent and unnecessary itemsets
- the generation of a large number of combinations that never occur in the database as well as
- the need to play out a few full database filters while creating continuous itemsets.

Proposed Research:-

In this paper, a Enhanced Apriori Algorithm is proposed to solve a major problem of the CAA using the principle of generating combinations from the frequent items found in each row of the transaction database.
Steps For the proposed work: -

- Collecting Data From Heterogeneous Database
- Minimum Support Threshold Definition module
- Frequent Items module
- Array of frequent items module
- Combination generation module
- Frequent itemsets module.

Collecting Data from Heterogeneous Database
Data integration is to provide a unified representation, storage and data management for various heterogeneous data environment, which is the basic function the heterogeneous data integration system must

implement. Information incorporation shields the heterogeneity of the different heterogeneous information sources, and completes brought together activity to various information sources through heterogeneous information reconciliation framework. Therefore, the integrated heterogeneous data is unified for users. The data forms involved in heterogeneous database are mainly structured data, semi-structured data and unstructured data three types. Structured data widely exists in a variety of information system database, the most common relational database. Semi-structured data commonly has Web pages as the chief representative, and XML can effectively manage and process such data. Unstructured data has common files, email and various documents. A practical information integration system should have intelligence, openness and initiative. Intelligence is to carry out unified processing, filtering, reduction, abstraction, integration and induction works for the structured, semi-structured and unstructured data from different databases .Openness is a heterogeneous and distributed database, which must solve the mismatching problem of the information expression with the structure. Initiative is to regulate the existing Internet data representation, exchange and service mechanism to provide proactive service mechanism

Minimum Support Threshold (MST)

Definition The MST is used in discovering frequent items and itemsets. The MST value is usually accepted as an input from the user since it is user defined. For this research however, the MST value was determined using Measures of Central Tendency. Measures of Central Tendency are statistical measures that make it possible to choose a value that best describes or represents a whole set of data. The MST for this research was determined using the measure known as the Mid-Range. The Mid-Range of a set of data is the mean of the maximum and minimum values in the dataset. The Mid-Range of the occurrence counts of the items in the transactions database was calculated and used as the MST. It is defined by:

M=(Max+Min)/2

M1=(Max+Min)/total transactions of filtered MST.

Frequent Items

Frequent items refer to those items with occurrence counts greater than or equal to the MST.

Array of Frequent Items

After the age of successive things, a void cluster with an indistinguishable structure from the Groceries table was made. The contents of the Groceries table were read into the array to populate it. The infrequent items were then deleted from the array.

Example:

I1	I2	I3	I4	I5
A	B	C		
A	C			
B	D	E		
A	C	F		
A	B	C	D	E
B	F			
B	C	E	F	
A	E	F		
A	B	C		
A	B			

OCCURRENCE COUNT OF ITEMS

ITEM	COUNT
A	7
B	7
C	6
D	2
E	4
F	4

Max occurrence is 7 and the Minimum occurrence is 2.

M= (7+2) / 2 = 4.5 ≈ 5 Therefore the MST is 5.

ITEM	COUNT
A	7
B	7
C	6

STRUCTURE OF ARRAY OF FREQUENT ITEMS

I1	I2	I3	I4	I5
A	B	C		
A	C			
B				
A	C			
A	B	C		
B				
B	C			
A				
A	B	C		
A	B			

Combination Generation

Starting with the first row to the last row in the array of frequent items, all possible combinations were generated using the items from that row.

Row1: AB, AC, BC

Row2: AC

Row3: -Row4: AC

Row5: AB, AC, BC

Row6: -Row7: BC

Row8: -Row9: AB, AC, BC

Row10: AB

ITEM	COUNT
AB	4
AC	5
BC	4

ITEM	Individual Performance %
AB	13.3
AC	16.6
BC	13.3

Combination Generation

Starting with the first row to the last row in the array of frequent items, all possible combinations were generated using the items from that row.

Frequent Itemsets

Frequent itemsets were obtained by counting the occurrence of each unique combination and checking if it was greater than or equal to the MST. All those combinations that met this criterion were selected as frequent itemsets.

IV Conclusion: -

In this research, an enhanced ENHANCED APRIORI ALGORITHM was developed to address a major problem of the CAA From the results, it was observed that the ENHANCED APRIORI ALGORITHM was better and more efficient than the CAA since it succeeded in eliminating the problem of non-existent combination generation.

V. REFERENCES:-

[1] Apachehadoop. http://hadoop.apache.org/, 2013.
[2] B. Goethals. Survey on frequent pattern mining. Univ. of Helsinki, 2003.
[3] Big Data Dimensions, http://www.klarity-analytics.com/392-dimensions-of-big-data.html
[4] Big Data Spectrum by Infosys, https://www.infosys.com/cloud/resource-center/.../big-data-spectrum.pdf
[5] C. C. Aggarwal and J. Han, Frequent Pattern Mining. Cham: Springer International Publishing, 2014.
[6] Dhruba Borthakur. The hadoop distributed file system: Architecture and design. Hadoop Project Website.
[7] J. Han, H. Pei, and Y. Yin. Mining Frequent Patterns without Candidate Generation. In: Proc. Conf. on the Management of Data (SIGMOD'00, Dallas, TX). ACM Press, New York, NY, USA 2000.
[8] K. Yu and J. Zhou. Parallel TID-based frequent pattern mining algorithm on a PC cluster and grid computing system. Expert Syst. Appl., vol. 37, no. 3, pp. 2486–2494, 2010.
[9] Kawuu W. Lin, Pei-Ling Chen, Weng-Long Chang. A novel frequent pattern mining algorithm for very large databases in cloud computing environments. In 2011 IEEE International Conference on Granular Computing (GrC), Page(s):399 –403.
[10] M. Riondato, J. A. De Brabant, R. Fonseca, and E. Upfal, PARMA: A parallel randomized algorithm for approximate association rules mining in MapReduce. in Proc. 21st ACM Int. Conf. Inf. Knowl. Manage., Maui, HI, USA, 2012, pp. 85–94.
[11] Manjit kaur, Urvashi Grag. ECLAT Algorithm for Frequent Itemsets Generation. International Journal of Computer Systems (ISSN: 2394-1065), Volume 01– Issue 03, December, 2014.
[12] P. Tan, M. Steinbach and V. Kumar, Introduction to data mining. Boston, Mass: Addison- Wesley, 2013.
[13] Rakesh Agrawal and Ramakrishnan Srikant, Fast algorithms for mining association rules in large databases. Proceedings of the 20th International Conference on Very Large Data Bases, VLDB, pages 487-499, Santiago, Chile, September 1994.
[14] Sandy Moens, Emin Aksehirli, and Bart Goethals. Frequent itemset mining for big data. In 2013 IEEE International Conference on Big Data, pages 111–118. IEEE, 2013.
[15] Sheela Gole and Bharat Tidke. Clustbigfim-Frequent Itemset Mining Of Big Data Using Pre-Processing Based On Mapreduce Framework. In International Journal in Foundations of Computer Science & Technology (IJFCST), Vol.5, No.3, May 2015.
[16] Y.-J. Tsay, T.-J. Hsu, and J.-R. Yu. FIUT: A new method for mining frequent itemsets. Inf. Sci.,vol. 179, no. 11, pp. 1724–1737, 2009.