A Comparative Study of Data Mining Techniques in Maintenance of Data warehouse

P.R.Vishwanath, Member, IEEE, Dr.D.Rajyalakshmi, Dr.M.Sreedhar Reddy

Abstract— Data warehouse plays a vital role in Decision support systems. As it needs to answer many complex queries, managerial level queries, requires advanced computing Techniques. Data mining is the process of extracting hidden useful information from huge data bases. Literature of Data Mining involving algorithms and techniques related to association, classification, clustering etc. Data warehouse deals with heterogeneous, huge amount of data and has to answer complex queries. Once Design Data warehouse is done then it need to concentrate on Data warehouse maintenance. Data warehouse requires advanced systems to increase the performance in addition to the existing system. The explicit-table approach, Materialized view (MV) approaches are to be added with latest techniques. Many methodologies for Optimization of calculating MV's and efficient calculating Data cubes are proposed. This paper explores the data mining techniques such as Association, clustering used for Data warehouse maintenance

Keywords— Data warehouse, Data mining, Data cubes, Maintenance, Clustering.

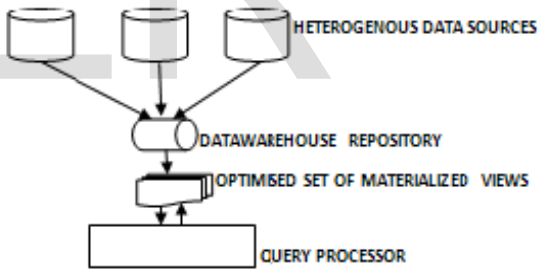
1. INTRODUCTION

According to W H Inman, A data warehouse is a subject-oriented, integrated, non volatile and Time-variant collection of data in support of management's decisions [1]. The tasks of data warehouse are data cleaning, data integration, data consolidation and summarization. Data warehouse uses the Update driven approach rather than query driven approach, in which data from multiple, heterogeneous sources is integrated in advance and stored in a warehouse/repository for direct querying and analysis. Different Data warehouse tools are used different Levels such as data modeling tools, ETL tools, Multi dimensional data base tools, reporting and analysis tools.

Data mining refers to extracting hidden useful knowledge from large amounts of data. The tasks of data mining are characterization, association, classification, prediction, clustering, outlier analysis etc...Data mining has become main tool in information technology, since huge amount data is converted into useful information and knowledge. Number of algorithms and approaches on each task on data mining are developed. These algorithms are used in different applications such as medical, clinical, banking, communication etc… These algorithms are also used in improving the performance of data warehouse (Data warehouse maintenance).

- P.R.vishwanath,dept of CSE ,RITS ,Chevella,+919985339968, email: vishwanath.raghava@gmail.com
- DrRajyalakshmi,dept of IT,GITAM,vizag,+9959728652 email: rdavuluri@gmail.com
- Dr.M.Sreedhar Reddy,Dept of CSE,SCET,Hyd,+9885619009 email: srircl@gmail.com

FIG 1: TYPICAL DATAWAREHOUSE MAINTANANCE



2. DATA WAREHOUSE MAINTENANCE PROBLEM

As the data warehouse services to supports large volumes of Data, It requires to have rapid and quick response to queries and to give accurate results. In data warehouse user query is not processed directly to access the source data (base tables), Intermediate repository is provided to access. The data sets generated in response to the query are called views .There views represents functions derived from base relations .To avoid re computing and selection of views ,we store some calculated results in central repository in order to improve the performance of data warehouse system. This system is known as materialization views. Materialized views are physical structures that improve data access time. The use of materialized views requires

additional storage space and entails maintenance overhead when refreshing the data warehouse. Select an appropriate set of materialized views (called a configuration of views), which minimizes total query response time and the cost of maintaining the selected views, given a limited storage space.

A Materialized View (MV) is the pre-calculated (materialized) result of a query. Unlike a simple VIEW the result of a Materialized View is stored in a table. Materialized Views are used when immediate response is needed and the query where the Materialized View bases on would take too long to produce a result. Materialized Views have to be refreshed for updating it once in a while. It depends on the requirements how often a Materialized View is refreshed and how actual its content is. Basically a Materialized View can be refreshed immediately or deferred; it can be refreshed fully or to a certain point in time. A MV is computed for SQL query since in a data warehouse, a materialized view relates to a SQL statement, that is materialized view corresponds to the result of SQL statement execution.

User requirements and constraints frequently changed. In order to accommodate novel and current requirements, there is a need of construction of innovative maintenance methods. Maintenance problem can be addressed through different parameters such as query selection cost, query maintenance cost, query storage space etc.

s3. ASSOCIATION MINING USED IN IMPROVING THE PERFORMANCE OF DATAWAREHOUSE

Association refers to mining frequent patterns, frequent items and correlations among the data in large transactional or relational data sets. Apriori algorithm-P Growth algorithms are developed to find some interesting frequent patterns.

T.Nalini, Dr. A.Kumaravel, Dr.K.Rangarajan [12] proposed An Efficient I-MINE Algorithm for Materialized Views in a Data Warehouse Environment. This work proposes cost effective mechanism for Materialized selection. Query frequency, query processing cost and space requirement are considered to materialize the candidate views. In this paper Data mining technique called I-Mine is used to generate frequently accessed queries. The advantage of the I-Mine algorithm is that it can mine the frequent queries with less computation time due to its I-Mine index structure compared with the traditional algorithms like, Apriori and FP-Growth [9].

Dr.T.Nalini, Dr.A.Kumaravel, Dr.K.Rangarajan[8] proposed a novel algorithm with im-lsi index for incremental maintenance of materialized view. In this algorithm Data mining technique called IM-LSI (Item set Mining using Latent Semantic Index) algorithm is used. This Item set Mining is part of Association Mining. This work materialize the candidate views by taking into consideration of query frequency, query processing cost and space requirement. To find the frequent queries this approach uses the F_P Growth algorithm, which is used for Item Set Mining. Then, an appropriate set of views can be selected to materialize by minimizing the total query response time and the storage space along with maximizing the query frequency [8].The outcome is used directly by the user to get quicker results for the queries. LSI(latent Semantic Index) is used to avoid re computation of MV,s when updating is done on the base table.

T.V.Vijay kumar, Kalyani Devi [22] proposed frequent queries identification for constructing materialized views. In this paper two data mining techniques association mining, clustering are used. In this paper selection of MV's are done based on past query pattern i.e. using the set of previously posed queries. This paper has four stages of implementation; they are Domain creation, Frequent queries identification, Optimal queries selection and Optimal queries merging. In this similar subject related queries are grouped together to form Domain creation. In this Hierarchical clustering is used to find clusters. In the merging process, Jaccard's co-efficient is used as the similarity measure. In the frequent queries identification the data mining technique called frequent mining is used. After the above four stages , the process remains with set of materialized views to be stored. This procedure reduces the time and ready to answer future queries.

Dong Xin, Jiawei Han, Xiaolei Li, Zheng Shao, and Benjamin W. Wah, [13] proposed Star-Cubing: Computing Iceberg Cubes by Top-Down and Bottom-Up Integration. To implement Data warehouse, one of the model proposed is Multi Dimensional Data model (MDM). To implement MDM we calculate the Data cubes. In this the important task is to calculate efficient computation of full or iceberg cubes(cubes that contain only aggregate cells whose measure value satisfies a threshold, called iceberg condition. In this paper the data mining technique called,

Apriori pruning is used to calculate Iceberg cubes. Apriori pruning is used to reduce unnecessary computation based on the anti-monotonic property, which is not possible in the top-down computation.

4. CLUSTERING TECHNIQUES USED IN SELECTION OF MATERIALIZED VIEWS

The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters [7]. Major clustering methods can be classified into four types. They are Partitioning methods, Hierarchical methods, Density-based methods, Model-based methods Grid-based methods. There are different algorithms developed from which some are k-means algorithm, k-medoids algorithm, agglomerative or divisive, BIRCH, DBSCAN and its extension, OPTICS, STING and COBWEB.

Kamel Aouiche, Pierre-Emmanuel Jouve, and Darmont [11] proposed Clustering based materialized selection using the clustering, one of the data mining technique. This algorithm uses workload approach. This procedure uses a query clustering involving similarity and dissimilarity measures defined on the workload queries, in order to capture the relationships existing between the candidate views derived from this workload. This makes use of query attribute matrix in the process of finding clusters .This algorithm builds the set of candidate views from the clusters formed. Further these candidate views are merged to resolve multiple queries. Since this merging process is applied on these clusters instead on all queries, reduces the computation cost. Thus the final clusters are materialized which reduces the space a lot.

Gang Zhao [10] proposed the CBDMVS algorithm (clustering based dynamic materialization view selection algorithm) which makes use of clustering technique to decrease the computational cost and space..In this method materialized views are clustered by using similarity function. Then these MV Clusters are dynamically adjusted. The algorithm finds the MV set which has relatively higher frequency responses performance to variety types of query. It uses the view replacement function. In this gain of every MV is calculated and stored. It eliminates the jitter. In this algorithm when updations are done only to the required materialized views but not whole MV set, which greatly reduces the computational cost.

Yogeshree D. Choudhari,Dr. S. K. Shrivastava [23] proposed the cluster based approach for selection of materialized views. The procedure uses the clustering of the views. This algorithm uses the record generator. Then System finds set of all possible queries resolved on generated records. Then Based on the access frequency set of queries are optimized. Further using the cluster area and threshold, the MV's are made. These are divided further into three types 1) Single query to Multi table MV. 2) Single query to single table MV. 3) Multiple queries to single table MV. This framework decreases the query response time.

5 COMPARATIVE STUDY

TABLE I COMPARISION OF VARIOUS RESEARCH WORKS

.	Data Mining Technique	Author	Algorithm Used	Approach
1.	Association Mining	T.Nalini Dr.Kumaravel Dr.K.Rangarajan	I-Mine Algorithm	Finding Frequent queries for reducing Materialized Views
2.	Association Mining	T.Nalini Dr.Kumaravel Dr.K.Rangarajan	I-Mine Algorithm Lsi	Finding Frequent queries for reducing Materialized Views and avoiding re computation of MV's by using LSi
3.	Association Mining	T.V.Vijay Kumar Kalyani Devi	Frequent Mining Clustering	Using Domain creation Clusters are formed and efficient merging process is adopted
4.	Association Mining	Dong Xin , Jiawei Han , Xiaolei Li , Zhang Shao, Benjamin . W.Wah	Apriori Algorithm	Apriori Purning is used to efficiently caluculate the iceberg cubes
5.	Clustering	Kamal Aouiche, Pierre Emmanuel Jouve , J'e'rome Darmont	Partition Based Clustering	Uses query attribute matrix there by using similarity, dissimilarity clusters are determined and used for storing Materialized Views
6.	Clustering	Gang Zhao	Partition Based Clustering	Uses similarity function, view replacement function to dynamic materialized view selection
7.	Clustering	Yogeshree. D Choudari , Dr.K.Shrivastava	Density Based Clustering	Framework which uses record generator, cluster area threshold to reduce the query response time

6 CONCLUSION

As the Data warehouse need to service the Decision support system, it needs to have efficient Query processing methods. Data warehouse query will not interact the repository directly, there by materialized views are pre calculated. It is impossible to construct all materialized views for any DW system. Many research proposals are developed for Optimization of the materialized views. Recent efficient approaches are using data mining

techniques to calculate MV's. This paper explored the Various Data mining techniques used for optimization of materialized views and calculating data cubes. Presently only Association, Classification techniques are widely used. This work may be extended to make use of classification, outlier analysis and advanced data mining algorithms to increase the performance of the Data warehouse process.

REFERENCES

[1] W. Inmon, "*Building the data warehouse*", Wiley publications, pp 23, 1991.

[2] R. Agrawal and R. Srikant, "*Fast Algorithm for Mining Association Rules*," Proc. 20th Int'l Conf. Very Large Data Bases (VLDB '94), Sept. 1994.

[3] R. Agrawal, T. Imilienski, and A. Swami, "*Mining Association Rules between Sets of Items in Large Databases*", Proc. ACM SIGMOD '93, May 1993.

[4] J. Han, J. Pei, and Y. Yin, "*Mining Frequent Patterns without Candidate Generation*", Proc. ACM SIGMOD, 2000.

[5] A. Savasere, E. Omiecinski, and S.B. Navathe, "*An Efficient Algorithm for Mining Association Rules in Large Databases*", Proc. 21st Int'l Conf. Very Large Data Bases (VLDB '95), pp. 432-444, 1995.

[6] V. Harinarayan, A. Rajaraman, and J. Ullman. "*Implementing data cubes efficiently*",Proceedings of ACM SIGMOD 1996 International Conference on Management of Data, Montreal, Canada, pages 205--216, 1996.

[7] Jiawei Han and Micheline Kamber , "*Data Mining Concepts and Techniques*, 2nd ed. Morgan Kaufmann Publishers, March 2006. ISBN 1-55860-901-6

[8] Nalini, T, Kumaravel, A and Rangarajan, K ,"*A Novel Algorithm With Im-Lsi Index For Incremental Maintenance Of Materialized Views*".Proceedings of JCS&T Vol. 12 No. 1, April 2012

[9] *Frequent Pattern Growth (FP-Growth) Algorithm An Introduction*,Florian Verhein ,January 2008

[10] An Gong, Weijing Zhao, "*Clustering-based Dynamic Materialized View Selection Algorithm*" Proceedings of Fifth International Conference on Fuzzy Systems and Knowledge Discovery, 2008, China,pp391-395

[11]K. Aouiche, P.Emmanuel Jouve, and J.Darmont ,"*Clustering-Based Materialized View Selection in Data Warehouses*"TechnicalReport, University of Lyon 2, 2007.

[12]T.Nalini,Dr.A.Kumaravel,Dr.K.Rangarajan ," *An Efficient I-Mine Algorithm For Materialized Views In A Data Warehouse Environment*", Ijcsi International Journal Of Computer ScienceIssues, Vol. 8, Issue 5, No 1, September 2011 Issn (Online):1694-0814

[13] Elena Baralis, Tania Cerquitelli, and Silvia Chiusano," *I-Mine: Index Support for Item Set Mining*" IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 4, april 2009

[14]D. Xin, J.Han, X. Li, and B.W.Wah. *Star-cubing: Computing iceberg cubes by top-downand bottom-up integration*. In Proc. 2003 Int. Conf. Very Large Data Bases (VLDB'03),Berlin, Germany, pages 476–487, Sept. 2003.

[15]J. Yang, and I. Chung, "*ASVMRT: Materialized view selection algorithm in data warehouse*", In International Journal of Information Processing System, 2006

[16] Gupta, H. & Mumick, I., *Selection of Views to Materialize in a Data Warehouse*. IEEETransactions on Knowledge and Data Engineering, 17(1), 24-43, 2005.

[17] A. Shukla, P. Deshpande, and J. F. Naughton, "*Materialized view selection for multidimensional datasets*," in Proc. 24th Int. Conf. V ery Large Data Bases, 1998, pp. 488–499.

[18] T.Nalini,S.K.Srivatsa,K.Rangarajan," International Journal of Advanced Research in Computer Engineering(IJARCE),"*Method of ranking in indexes on materialized view for database workload*" Vol.4,No.1,pp 157-162

[19]A. N. M. Bazlur Rashid and M. S. Islam, *"An Incremental View Materialization Approach in ORDBMS,"* IEEE Intl. Conf. on Recent Trends in Information, Telecommunication and Computing 2009.

[20]J.Yang, K. Karlapalem, and Q.Li. "*A framework for designing materialized views in data warehousing environment*". Proceedings of 17th IEEE International conference on Distributed Computing Systems, Maryland, U.S.A., May 1997.

[21]H. Gupta. "*Selection of Views to Materialize in a Data Warehouse".* Proceedings of International Conference on Database Theory, Athens, Greece 1997.

[22] T.V.Vijay kumar and Kalyani Devi,"*Frequent queries identification for constructing materialized views",*Journal of Computer Science & Technology (JCS&T);2012, Vol. 12 Issue 1, p32.

[23]Yogeshree D. Choudhari,Dr. S. K. Shrivastava "*Cluster Based Approach for Selection of Materialized Views* " proceedings of International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE)july2012

[24] Y. Zhuge, H. Garcia-Molina, J. Hammer, and J. Widom, "*View Maintenance in a Warehousing Environment*." In Proceedings of the ACM SIGMOD Conference, San Jose, California, May 1995.

[25] M. El-Hajj and O.R. Zaiane, "*Inverted Matrix: Efficient Discovery of Frequent Items in Large Datasets in the Context of Interactive Mining*," Proc. Ninth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD), 2003.s

Viswanath Raghava .P born in Hyderabad completed B.Tech and M.Tech from JNTU is currently working as Associate professor at Royal Institute of technology and Science, Hyderabad. He is presentlydoing research in Data mining and data warehousing at GitamUniversity, Vizag. He has published 6 international and1national journals.

Dr. D.RajyaLakshmi is working as Professor in the department of Information Technology, GITAM Institute of Technology, GITAM University, Andhra Pradesh, India. She was awarded PhD from Jawaharlal Nehru Techn ological University in the area of Image Processing. She has 18 years of teaching and researchexperience. She has more than twenty six papers were presented and published in various conferences and journals respectively.

Dr. M. Sreedhar Reddy is currently working as professor and Head of Geetanjali of Engineering and Technology, institute, JNTU, Hyderabad. He was awarded PhD from Nagarjuna University He has published 15 international and nine national level conferences and journals. His research areas software engineerissng, Data mining, Networks and E-commerce.