Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 2, Issue. 12, December 2013, pg.318 – 321

REVIEW ARTICLE

Literature Review on Information Extraction by Partitioning

Sonam S.Chauhan[1], Prashant R. Deshmukh
[1]Department of Information Technology
Sipna College of Engineering and Technology,
SANT GADGE BABA AMRAVATI UNIVERSITY, AMRAVATI
[1]Email:sonams.chauhan@yahoo.in

***Abstract*— Information extraction systems are implemented traditionally as a pipeline of special-purpose processing modules that target at the extraction of a specific type of information. Such an approach has major drawback that whenever a module is improved or a new extraction goal emerges, extraction need to be applied to all parts even if only a small part is affected.**

Here we describe an approach for information extraction in which we partition the dimensions (attributes), i.e., a higher dimension of large data set can be transformed into relatively smaller subsets of data in certain numbers that might be processed easily. Here we use hetrogeneous database as large data set. Thereafter, based on the separation of dimensions, the discernible dataset of all data are computed so as to get their core attribute sets. Furthermore, the attribute reduction and data redundancy methods are used to obtain the partition results.

In our approach towards information extraction we tried to partition are data sets so that information can be extracted easily and correctly with query processing which generate results in less time. Feasibility of our incremental extraction approach can be demonstrated by performing experiments to spotlight two important aspects of an information extraction system: quality of extraction results and efficiency.

***Keywords*— Hetrogeneous databases; multidimensional data; partitioning; association rules; algorithm; itemsets**

I. INTRODUCTION

In the recent years our capabilities of both collecting and generating data have been grown rapidly. In business management, government administration, scientific and engineering data management, and much other application millions of databases have been used. It has been noted that the number of such databases is growing rapidly as a result of the availability of affordable and powerful database system. This rapid growth in databases and data has resulted in urgent need for new tools and techniques that can rapidly, intelligently and automatically transform the processed data into more useful information and knowledge. Consequently, data extraction has become a research area with increasing importance [1, 2, 3]. By knowledge discovery in database, interesting knowledge, regularities, or high-level information can be extracted from the relevant sets of data in database and can be investigated from different angles and large databases thereby serve as rich and reliable sources for knowledge generation and verification.

© 2013, IJCSMC All Rights Reserved

Knowledge and Information from huge databases has been reorganized as a key research topic by many researchers database system. The discovered knowledge can be applied to information management, query processing, decision making, process control and much application. Furthermore several emerging application in information providing services, also call for various data extraction techniques to better understand user behaviour, to meliorate the service provided, and to increase opportunities. In response to such a demand, here we describe an approach for information extraction in which we partition the dimensions (attributes), i.e., a large data set having a higher dimension can be transformed into certain number of relatively smaller data subsets that can be easily processed. Here we use hetrogeneous database as large data set. Then, based on the partitioning of dimensions, the discernible dataset of all data are computed to obtain their core attribute sets. Furthermore, the attribute reduction and data redundancy methods are used to obtain the partition results. In our approach towards information extraction we are trying to design such an algorithm which will be a generalized version of Apriori algorithm that dynamically generates itemset by decreasing number of passes [4].

II. PREVIOUS WORK

Extracting knowledge and information from huge databases has been recognized as a key research topic by many researchers in database system and machine learning and by large number of industrial companies as an important area with an opportunity of major revenues. Many papers and techniques have been proposed for providing solutions in data extraction from bulky data stored. The major source of our inspiration is early work by Jin Zhou, Liang Hu, Feng Wang, Huimin Lu, and Kuo Zhao in [5]. In this paper, an efficient multidimensional fusion algorithm is proposed for IoT data that is based on partitioning. Partitioning is the basic concept involved here. In rough set theory the discernible matrixes of all data subsets are computed based on the partitioning of dimensions to obtain their core attribute sets. Moreover, a global core attribute set can be determined. Finally, fusion result is obtained by the rule extraction and attributes reduction methods. By means of proving a simulation and a few theorems, the effectiveness and correctness of this algorithm is illustrated.

Many caching mechanism was also described in many papers which could extract information from bulky data in less time, however many times threatened the goal of quality of extraction result.

We also referred to some papers describing the algorithms for association rules like Apriori algorithm [4]. However it was discovered that the step of determining itemsets by scanning the whole database and testing each transaction against the hash tree build is very expensive.

III. PROPOSED APPROACH

In order to determine an acceptable quality of intelligence, diverse and voluminous data have to be combined, integrated and partitioned. In this paper, we propose an efficient computational method for large data. The advantage of this method is that it transforms a big problem into smaller problems that are easier to process; this substantially reduces the temporal and spatial computational complexities. Several techniques have been presented in order to illustrate the correctness of the proposed approach. Here we describe an approach for information extraction in which we partition the dimensions (attributes), i.e., higher dimensions of large data set can be transformed into relatively smaller data subsets in certain numbers that can be easily processed. Here

© 2013, IJCSMC All Rights Reserved

we use hetrogeneous database as large data set. Then, based on the partitioning of dimensions, the discernible dataset of all data are computed to obtain their core attribute sets. Furthermore, the attribute reduction and data redundancy methods are used to obtain the partition results. An example of our partition approach is as below. The data in different database is partitioned and brought in Partition database.

Oracle DB

e-id	Name	Address	P.No	Salary	Post
111	Ram	Adc,pune	00025678	40000	Manager
112	Ajay	Pqr,goa	00025878	10000	Clerk
113	Kale	Abc,Delhi	00027678	10000	Clerk
114	Rita	Pqr,goa	00026678	10000	Clerk

SQL DB

e-id	Name	Post	Salary
01	Ajay	Manager	40000
02	Kale	Clerk	20000
03	Riya	Clerk	20000

Access DB

e-id	Name	Post	Salary
001	Ram	Manager	40000
002	Ajay	Clerk	10000
003	Kale	Clerk	10000

Fig. 1: Data in Hetrogeneous Database

Partition DB

Part 1

e-id	Name	Salary	Database
111	Ram	10000	Oracle
112	Ajay	10000	Oracle
113	Kale	10000	Oracle
114	Rita	20000	Oracle
01	Ram	40000	Access
02	Ajay	20000	Access
03	Riya	20000	Access
001	Ram	10000	SQL
002	Ajay	10000	SQL
003	Kale	10000	SQL

Part 2

e-id	Post
111	Manager
112	Clerk
113	Clerk
114	Clerk
01	Manager
02	Clerk
03	Clerk
001	Manager
002	Clerk
003	Clerk

Fig. 2: Data in Partition Database

© 2013, IJCSMC All Rights Reserved

Here we will describe an algorithm to generate large itemsets that uses only 2 passes to generate large itemsets which effectively reduces the estimation cost and also reduce the size of Transaction Database. Here instead of keeping a record of counting the occurrence of item in number of transaction we will compute the percentage of occurrence of an item in fixed number of transaction and filter out most of the unused items from the partition database. This will not only reduce the data base size but also reduces the estimation cost. Our approach towards information extraction will not only perform query processing that generate result in less time but will also focus on Quality extraction.

IV. CONCLUSION

The Partition Approach proposed by us can be used to extract data from any Hetrogeneous databases in less time. However it is more suitable to databases were updating of database is less as large updating results in frequent updating in Partition database, increment in estimation cost. The advantage of this approach is that it transforms a big problem into smaller problems that are easier to process; this substantially reduces the temporal and spatial computational complexities. To demonstrate the feasibility of our incremental extraction approach, we performed experiments to highlight two important aspects of an information extraction system: efficiency and quality of extraction results. Further, we perform a simulation to enumerate the better efficiency and effectiveness of the proposed work. In a future study, the results of the partition data will be presented. The relationships between the number of dimensions, number of partitions, and volume of objects and their influence on the computation efficiency will be discussed. Furthermore, the effect of these relationships on the partitions will be investigated.

REFERENCES

[1] U.M. Fayyad, G.Piatetsky-shapiro, P.Smyth, and R. Uthurusamy, Advances in knowledge Discovery and Data Mining, Mining, AAAI 1996.

[2] G.Piatetsky-shapiro Advances in knowledge Discovery and Data Mining, AAAI/MIT 1991.

[3] A.Sliberschatz, M.Stonebraker, and J.D.Ullman,"database Research: Achievement and opportunities into 21st Century,"May1995.

[4] Rakesh Agrawal and Ramakrishnan Srikant: Fast Algorithm for Mining Association Rules.1994.

[5] Jin Zhou, Liang Hu, Feng Wang, Huimin Lu, and Kuo Zhao in "An Efficient Multidimensional Fusion Algorithm for IoT Data Based on Partitioning" Aug 2011.

[6] H. Patni, C. Henson, and A. Sheth, Linked sensor data in Proc. 2010.

[7] A. K. Joshi, P. Jain, P. Hitzler, P. Z. Yeh, K. Verma, A.P. Sheth, and M. Damova, Alignment-based querying of linked open data, in Proc. on the Move to Meaning Internet Systems: OTM 2012, Confederated International Conferences: Coop IS, DOA-SVI, and ODBASE 2012, Rome, Italy, 2012, pp. 1-18.

[8] A. Rajaraman, J. Leskovec, and J. D. Ullman, Mining of Massive Datasets. Cambridge University Press, 2010.

[9] N. Pham and R. Pagh, A near-linear time approximation algorithm for angle-based outlier detection in high dimensional Data, in Proc. 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'12), Beijing, China, 2012, pp. 877-885.

[10] S. Kahramanli, M. Hacibeyoglu, and A. Arslan, Attribute reduction by partitioning the minimized discernibility function, 2011.

© 2013, IJCSMC All Rights Reserved