A Clustering OLAP Analysis in a Big Data Stream Environment

Ho Jin Woo¹, Kil Hong Joo², Nam Hun Park³

¹ Dept. Of Computer Science, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul, Korea, judas@database.yonsei.ac.kr
² Dept. Of Computer Education, Gyeongin National University of Education, San 6-8 Seoksu-dong Manangu Anyangsi, Gyeonggi, Korea, 430-040 khjoo@ginue.ac.kr
³ Dept. Of Computer Science, Anyang University, 102 Samsungli, Buleunmyun, Ganghwagun, Incheon, Korea, 417-833 nmhnpark@anyang.ac.kr

Abstract. The online analytical processing (OLAP) is a process where an end user directly accesses multi-dimensional information to analyze information in a dialogue type for his or her decision making [1]. OLAP is a multi-dimensional data modeling technique utilizing data cube and data cube is consisted of the measured values subject to analysis and dimensions serving as analysis perspective [3]. Each dimension becomes the axis of the cube and the measured values are included in cells produced by the combinations of attribute values structuring dimension and then saved. Cells produced in this manner are not only huge in their number but also require high cost to count for the measured value saved in each cell. Especially, every time data are added, the whole data cube should be renewed by repeating the counting calculation. So it is limited to apply the method to the recent big data environment including smart phones and SNS. To address such limitations, this research combines the existing data cube modeling with clustering structure to perform multi-dimensional big data stream OLAP analysis. The suggested system groups each dimensional attribute values that

Keywords: Big data, Data streams, OLAP, Clustering

1 Introduction

The online analytical processing (OLAP) is a process where an end user directly accesses multi-dimensional information to analyze information in a dialogue type for his or her decision making [1]. OLAP is a multi-dimensional data modeling technique utilizing data cube and data cube is consisted of the measured values subject to analysis and dimensions serving as analysis perspective [3]. Each dimension becomes the axis of the cube and the measured values are included in cells produced by the combinations of attribute values structuring dimension and then saved. Cells produced in this manner are not only huge in their number but also require high cost to count for the measured value saved in each cell. Especially, every time data are added, the whole data cube should be renewed by repeating the counting calculation. So it is limited to apply the method to the recent big data environment including smart phones and SNS. To address such limitations, this research combines the existing data cube modeling with clustering structure to perform multi-dimensional big data stream OLAP analysis. The suggested system groups each dimensional attribute values that
frequently occur in data and save them as cluster information. It also adds a structure maintaining the attribute values of measured values for easier data analysis.

2 Clustering-based OLAP analysis

The proposed method bases sub-space clustering [4] to produce frequently-occurring data sets as frequent cuboid. The figure 1 shows an example where each tuple is formed as frequent cuboid in the data stream with the schema of <Company, Region, Color, Sales>. The information of tuple in data stream with <Company, Region, Color> being {BMW, America, White}, {BMW, Asia, White} is included in the first frequent cuboid \(C_{11}, C_{21}, C_{31}\). The tuple information of data stream with {GM, Europe, Black}, {VW, Europe, Red} is included in the second frequent cuboid \(C_{12}, C_{22}, C_{32}\). {Ford, Asia, White} was excluded from frequent cuboid as its frequency was low. By using such frequent cuboids, different sub-data cubes can be generated according to the combinations formed under cuboid data sets. It is defined as frequent sub data cube. Frequent sub data cube is consisted of the data set forming frequent cuboid and their measured values. The frequent sub data cube of the frequent cuboid in the figure 1 is as follows: The data sets for \(C_{11}, C_{21}, C_{31}\) are {BMW, America, White}, {BMW, Asia, White}. The data sets for \(C_{12}, C_{22}, C_{32}\) are {GM, Europe, Black}, {GM, Europe, Red}, {VW, Europe, Black}, {VW, Europe, Red}.

The process of calculating the measured values by using the sub data cube for specific inquiry conditions such as, for example, Company, Region, Color being 2, 1 and 2. For the node \(c_{11}\) including Company = 2, standard normal distribution \(Z_{\text{Company}=2}=(2-c_{11}\mu)/c_{11}\sigma\) was utilized to calculate the probability of the distribution of value 2 in the cluster. Then by multiplying the number by \(c_{11}\)'s measured value of 100 to calculate the measured value of 34.13 when \{Company\} = \{2\}. Also, for node \(c_{21}\) including Region = 1, standard normal distribution \(Z_{\text{Region}=1}=(1-c_{21}\mu)/c_{21}\sigma\) is used to calculate the probability of the value 1’s distribution in the cluster. By multiplying the number by 34.13, the measured value when \{Company\} = \{2\} to produce the value of 11.65 when \{Company, Region\} = \{2, 1\}. Lastly, for \(c_{31}\) including Color = 2, \(Z_{\text{Color}=2}=(2-c_{31}\mu)/c_{31}\sigma\) is used to calculate the probability of value 2’s distribution in the cluster.

3 Experimental evaluations

The clustering OLAP analysis results were examined by utilizing various data produced according to Zipfian distribution that reflects actual data distribution. Zipfian distribution between 1.8~3.4 were utilized as data sets. There are 4~8 dimensional attributes and 100,000 tuples. Each dimensional attribute values is consisted of 100 mutually different values. The figure 2 shows the performance comparisons among Full Data Cube (FDC), Full Stream Cube (FSC) [2], and Dynamic Cube Tree (DCT). In order to compare with other cubes, the stream cube has expanded its two layers to the full cube. The figure 2-(a)
shows the memory usage of each data cube while the distribution of data is getting crowded to one side due to the increase of the Zipf distribution. The dynamic cube tree uses less memory space if $S_{\text{min}}, S_{\text{max}}$, and Zipf distribution have higher values. The figure 4-(b) shows the memory usage while increasing the number of tuples. The dynamic cube tree shows stable memory usage for the same distribution, but the other two cubes show an increase of memory usage because they try to store the newly generated tuples which they didn’t have before.

![Figure 1: Composition of frequent cuboid](image1.png)

![Figure 2: The performance comparisons among three cubes](image2.png)
4 Conclusion

In this research, an OLAP analysis technique was proposed based on the clustering of big data stream where data distribution characteristics change rapidly. Existing methods form nodes for every attribute value so they consume larger memory as the number of dimensional attributes rises. On the other hand, the suggested method introduced clustering into the data cube modeling method to collect information of clusters where each attribute value belong and form frequent cuboid tree nodes. The method, thought its precision is slightly compromised, was found to use less memory and shorter time to process. Based on these findings, the proposed method is expected to function usefully in OLAP analysis under the big data stream environment.

Acknowledgements. This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science Technology(NRF-2014R1A1A1008524).

References

2. J. Han, Y. Chen, G. Dong, J. Pei, B. W. Wah, J. Wang and Y. D. Cai, Stream Cube: An Architecture for Multi-Dimensional Analysis of Data Streams, Distributed and Parallel Databases, 18, 2 (2005)