32

Horizontal Fragmentation of Data Warehouses Using Decision Trees

Nidia Rodríguez Mazahua, Lisbeth Rodríguez Mazahua, Asdrúbal López Chau and Giner Alor Hernández

Horizontal Fragmentation of Data Warehouses Using Decision Trees

Nidia Rodríguez Mazahua[1], Lisbeth Rodríguez Mazahua[1], Asdrúbal López Chau[2] and Giner Alor Hernández[1]

[1]Tecnológico Nacional de México/ IT Orizaba, División de Estudios de Posgrado e Investigación. Orizaba, Veracruz, México
[2]Universidad Autónoma del Estado de México, Centro Universitario UAEM Zumpango, Estado de México, México
nydirodriguez1@gmail.com, lrodriguezm@ito-depi.edu.mx, alchau@uaemex.mx, galor@ito-depi.edu.mx

Abstract. One of the main problems faced by Data Warehouse (DW) designers is fragmentation. Several studies have proposed data mining-based horizontal fragmentation methods, which focus on optimizing the query response time and execution cost to make the DW more efficient. However, to the best of our knowledge there not exist a horizontal fragmentation technique that uses a decision tree to carry out fragmentation. Given the importance of decision trees in classification, since they allow obtaining pure partitions (subsets of tuples) in a data set using measures such as Information Gain, Gain Ratio and the Gini Index, the aim of this work is to use decision trees in the DW fragmentation. For this, the requirements necessary to carry out horizontal fragmentation using decision trees will be determined, and the fragmentation method will be designed, which will consist of determining the most frequent OLAP (On-line Analytical Processing) queries, analyzing the predicates used by the queries, and based on this build the decision tree, from which the horizontal fragments will be generated. The method will be implemented and validated using a case study in tourism.

Keywords: CLQ, Fragmentation, Data warehouse, Decision Trees.

1 Introduction

A Data Warehouse is a theme-oriented, integrated, time variable, and non-volatile data collection in support of management's decision-making process. Data Warehousing provides architectures and tools for business executives to systematically organize, understand, and use the data to make strategic decisions [1].

On the other hand, fragmentation is a distributed database design technique that consists of dividing each database relation in smaller fragments and treating each fragment as an object in the database separately, there are three alternatives for that: horizontal, vertical, and hybrid fragmentation [2].

Decision tree classifiers are so popular because their construction does not require any domain knowledge or parameter setting, they can handle multidimensional data, the learning and classification steps of decision tree induction are simple and fast, and they have good accuracy [1]. Since horizontal fragmentation consists of dividing a database table into subsets of tuples and decision trees have been used to horizontally fragment relational database tables based on OLTP (On-line Transaction Processing) workloads [3], in this paper, we propose using a decision tree to horizontally fragment a Data Warehouse. This paper is made up of the following parts: Section 2 go through the related works on DW horizontal fragmentation, Section 3 set the research objective of work, in Section 4 the stages that conform the methodology to be followed are described, Section 5 reports the preliminary achievements in the work, and finally, the article is concluded in Section 6.

2 Related work

The approach proposed in [5] consisted of an incremental horizontal fragmentation technique of the DW through a web service. In [6], it was proposed a method called Chabok, which uses two-phase Map Reduce to solve DW problems with big data. As part of their ongoing work on workload–driven partitioning, Boissier and Kurzynski [7] implemented an approach called aggressive data skipping and extended it to handle both analytical and transactional access patterns. In [8] it was used linear programming to solve the NP-hard problem of determining a horizontal fragmentation scheme in relational DW. Also in [9], the authors proposed a graph-based database partitioning method called GPT that improves the performance of queries with less data redundancy. In [10] it was proposed a dynamic fragmentation strategy for OLAP cubes using association rule mining. Finally, in [11] it was presented a horizontal data partitioning approach tailored to a large DW, interrogated through a high number of queries, the idea was to fragment horizontally only the large fact table based on partitioning predicates, elected from the set of selection predicates used by analytic queries.

3 Research objective

The main objective of this work is to develop a horizontal fragmentation method that uses decision trees to obtain fragmentation schemes that achieve query optimization in a DW.

4 Methodology

The process to develop the fragmentation method which uses decision trees consists of four stages:

1. **Analysis.** In this stage, it is necessary to realize a comparative analysis of horizontal fragmentation methods and their cost models to know their advantages and disadvantages and select the most suitable for a comparison with the proposed method and cost model that will be performed in the evaluation stage.
2. **Design**. The decision tree-based horizontal fragmentation method for DW and the cost model to compare different horizontal fragmentation schemes will be designed in this stage, moreover, we will develop a tourism DW following the Kimball [12] methodology. This tourism DW will be useful for the evaluation stage. The tourism data will be integrated into the DW from different sources such as SECTUR (Secretaría de Turismo) and INEGI (Instituto Nacional de Estadística y Geografía).
3. **Implementation.** The proposed fragmentation method and cost model will be implemented in two data warehouses: the tourism DW and the SSB (Star Schema Benchmark).
4. **Evaluation.** The proposed method will be compared with the selected techniques of the comparative analysis to demonstrate the strengths of our method.

5 Achievements

The state of the art concerning to the horizontal fragmentation methods were analyzed to know their advantages and disadvantages, also the cost models used by methods was compared to find the most relevant costs considered by the authors when obtaining their fragmentation scheme in DW. 39 methods and 24 cost models formed the analysis. The horizontal fragmentation methods were analyzed by 6 criteria finding the following results: 1) Type of scheme used: the star scheme was the most multidimensional model used by methods; 2) Completeness: 34 methods presented the necessary information to implement them; 3) Validation: 30 techniques used benchmarks for their evaluation, the most used benchmarks were APB-1, APB-1 Release II, and TPC-H; 4) Implementation easiness: more than half of methods are classified as difficult to implement mainly because some of them focus on complex DW fragmentation, while others perform incremental (dynamic) fragmentation; 5) Cost model: 24 methods included a cost model; 6) Type and database management system used: the data model most used was the relational one. The most used DBMS was Oracle 11g. Whit respect to cost models, the criteria analyzed were costs considered and if they corresponded to local processing and/or communication. Almost all cost models only take into account the local processing cost, only [13] also considers the communication cost in its model.

6 Expected results

The expected results for this work are mainly to develop a horizontal fragmentation method that uses decision trees to obtain a partitioning scheme which optimizes the response time and the execution cost of the OLAP queries in a DW, to validate the method it will be applied in a case study of tourism in México.

7 Conclusions

In this work, a horizontal fragmentation method is being developed to take advantage of the potential of decision trees in the classification to adapt them in the process of DW horizontal fragmentation. Until now, interesting results have been found which are leading the investigation.

References

[1] J. Han, M. Kamber, y J. Pei, *Data Mining Concepts and Techniques*. USA: Morgan Kaufmann Publishers., 2012.

[2] M. T. Ozsu y P. Valduriez, *Principles of Distributed Database Systems*, Fourth edition. Cham, Switzerland: Springer Nature Switzerland AG, 2020.

[3] C. Curino, E. Jones, Y. Zhang, y S. Madden, "Schism: a workload-driven approach to database replication and partitioning", *Proc. VLDB Endow.*, vol. 3, núm. 1–2, pp. 48–57, sep. 2010, doi: 10.14778/1920841.1920853.

[4] R. C. Mateus, T. L. L. Siqueira, V. C. Times, R. R. Ciferri, y C. D. de Aguiar Ciferri, "Spatial data warehouses and spatial OLAP come towards the cloud: design and performance", *Distrib. Parallel Databases*, vol. 34, núm. 3, pp. 425–461, sep. 2016.

[5] E. Abdelaziz y M. Ouzzif, "Web Service for Incremental and Automatic Data Warehouses Fragmentation", *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, núm. 6, pp. 1–10, 2017.

[6] M. Barkhordari y M. Niamanesh, "Chabok: a Map-Reduce based method to solve data warehouse problems", *J. Big Data*, vol. 5, núm. 40, pp. 1–25, dic. 2018.

[7] M. Boissier y K. Daniel, "Workload-Driven Horizontal Partitioning and Pruning for Large HTAP Systems", en *2018 IEEE 34th International Conference on Data Engineering Workshops (ICDEW)*, Paris, abr. 2018, pp. 116–121.

[8] M. Barr, K. Boukhalfa, y K. Bouibede, "Bi-Objective Optimization Method for Horizontal Fragmentation Problem in Relational Data Warehouses as a Linear Programming Problem", *Appl. Artif. Intell.*, vol. 32, núm. 9–10, pp. 907–923, nov. 2018.

[9] Y.-M. Nam, M.-S. Kim, y D. Han, "A Graph-Based Database Partitioning Method for Parallel OLAP Query Processing", en *2018 IEEE 34th International Conference on Data Engineering (ICDE)*, Paris, abr. 2018, pp. 1025–1036.

[10] K. Letrache, O. El Beggar, y M. Ramdani, "OLAP cube partitioning based on association rules method", *Appl. Intell.*, vol. 49, núm. 2, pp. 420–434, feb. 2019.

[11] M. Kechar y S. N. Bahloul, "Bringing Together Physical Design and Fast Querying of Large Data Warehouses: A New Data Partitioning Strategy", presentado en BDIoT'19: Proc. of the 4th Int. Conf. on Big Data and Internet of Things, Rabat Morocco, 2019.

[12] R. Kimball y M. Ross, *The Kimball Group Reader: Relentlessly Practical Tools for Data Warehousing and Business Intelligence*. Indianapolis, Indiana: John Wiley & Sons, Inc., 2016.

[13] Pedro Furtado, "Experimental Evidence on Partitioning in Parallel Data Warehouses", *Proc. 7th ACM Int. Workshop Data Warehous. OLAP*, pp. 23–30, 2004.