Engineering Applications of Artificial Intelligence 71 (2018) 125–137

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

A game theory based framework for materialized view selection in data warehouses

Hossein Azgomi, Mohammad Karim Sohrabi *

Department of Computer Engineering, Semnan Branch, Islamic Azad University, Semnan, Iran

ARTICLE INFO

Keywords:
Materialized view selection
Multi-value processing plan
Data warehouse
On-line analytical processing
Game theory

ABSTRACT

Data warehouses exploit On-Line Analytical Processing (OLAP) to make rapid answers for analytical queries. Huge amount of aggregated data within a data warehouse on the one hand, and complex analytical queries raised in a data warehouse on the other hand, increase response time to queries extremely. To solve this problem, a number of views are derived and extracted from original base tables and queries have been answered using them. Since materialization of all possible views is not effective because of limitation of storage and maintenance overhead, selecting an optimal set of views for materialization is crucial to maximize data warehouse performance.

In this paper, a game theory based framework for the materialized view selection is proposed. In the proposed framework, query processing and view maintenance costs play a game against each other as two players and continue the game until reach the equilibrium. According to the framework, a new static method, called Game Theory based Materialized View selection (GTMV), has been proposed. Verification of proposed approach has been evaluated using several synthetic and real world datasets. Experimental results show that the GTMV method has better performance comparing previous algorithms and substantially outperform former methods.

1. Introduction

Great importance of data analysis in today's information based and knowledge based world is evident. On-Line Analytical Processing (OLAP) is one of the involved technologies for analyzing data which is employed by data warehouses (Mansmann et al., 2014). Data warehouse is an integrated repository of information collected from several operational databases, with some important characteristics such as subject-orientation, integrity, time variance, and non volatility, which is used to obtain analytical and statistical information for decision support systems (Lechtenbörger and Vossen, 2003). Massive amount of data and complex queries lead to unacceptable increase of response time and emphasize needing to create an efficient mechanism to rapid response to such queries. Materialized view selection is the most important and the most applicable solution. An optimal set of views, according to some constraints such as memory space limitation and maintenance cost, are selected and materialized in this approach. Materialized views are employed to efficient processing of complex queries. Finding this optimal set of views is a challenging issue that has attracted a lot of recent studies (Huang et al., 2014).

From a general point of view, materialized view selection methods can be divided into two categories: static view selection and dynamic view selection. In static view selection methods, selecting and materializing of views take place before executing the first query and the set of materialized view does not change until the end of execution of last query. In dynamic view selection methods, the set of materialized views changes during execution of queries, and some of the selected views have been removed and are replaced by other ones. The dynamic methods are also used in two ways. In the first type, the set of all queries and their execution order are known in advance, and in the second type, queries are not known from the beginning and added to the workload gradually (Mami and Bellahsene, 2012). Dynamic algorithms are more successful than static ones and most of efficient materialized view selection methods exploit dynamic algorithms to find the optimal set of appropriate views.

Most of materialized view selection methods organize the set of all views as a unified structure firstly, and then select the set of optimal view for materialization using this structure. Multiple View Processing Plan (MVPP) (Yang et al., 1997), AND–OR DAG (Roy et al., 2000), and Data Cube Lattice (Harinarayan et al., 1996) are the most well known applicable types of these structures. MVPP is a directed acyclic graph, in which, roots are queries, leaves are base relations, and other nodes are relational operators, such as selection, projection, join, and data

* Corresponding author.
E-mail addresses: Hossein.Azgomi@gmail.com (H. Azgomi), Amir_sohraby@aut.ac.ir (M.K. Sohrabi).

https://doi.org/10.1016/j.engappai.2018.02.018
Received 27 May 2017; Received in revised form 9 January 2018; Accepted 25 February 2018
Available online 20 March 2018

aggregation, which are used to write a query. In fact, query processing plan of multiple queries are combined as the MVPP. AND–OR DAG is also a directed acyclic graph which is composed of two types of nodes, namely operation nodes and equivalent nodes. Each operation node contains a relational algebra expression, and each equivalent node displays a set of equivalent logical expressions. Root of AND–OR graph is the result of the query execution and its leaves are base relations. Equivalent nodes are the main difference between MVPP and AND–OR DAG. In data cube lattice, each node shows a view or a query. Edges of the lattice demonstrate the relationship between the views. In this structure, a query can be answered and update using other queries. There also exists some strategies in which the process of materialized view selection is done without a particular structure. Some methods have used query rewriting to modify queries so that more common views can be obtained from them and they could exploit each other's results (Halevy, 2001). Furthermore, there exist some methods which select the optimal views using syntactical analysis of the workload. These methods investigate and evaluate the workload and select a subset of base relations such that for them and their corresponding selected views for materialization, the cost of the workload is significantly reduced. In this case, the search space for finding the set of optimal views is very large (Mami and Bellahsene, 2012).

In term of limitations, methods are provided to solving the problem of materialized view selection can be divided into different categories. Some methods select the appropriate set of view for materialization without any limitation. These views are selected so that the costs of query processing and view maintenance are minimal. In some methods, view selection has been carried out under storage space restriction. A specified given memory space is supposed in these methods and an optimal set of selected views, so that the summation of their required storage space to materialization is less than or equal to the available memory space, is desired. In some other methods, the view selection problem should satisfy the view maintenance cost constraint. In this category of methods, a limited time is given, in which all selected materialized views should be updated if a subset of base relations has been modified. There exist a few materialized view selection algorithms which consider both memory space and view maintenance cost limitations.

Games theory is a subset of mathematics that attempts to use the design and analysis of scenarios to predict behaviors and decision results of autonomous interacting entities (Nisan et al., 2007). Game theory is used in various disciplines, including computer science (Shoham, 2008). Distributed computing (Grosu and Chronopoulos, 2005), game-based learning (Tobias et al., 2014), Networks' routing (Vallet et al., 2016), network security (Liang and Xiao, 2013), decision support systems (Carreras et al., 2011), and data mining (Wang, 2006) are some of the applications of game theory in the field of computer science. In data related issues, game theory can be used in two ways. On the one hand, data related problems, such as extracting hidden patterns from a dataset, can be solved using game theory techniques, and on the other hand, with the help of data mining and knowledge discovery techniques, game theory can provide useful information for solving various problems. As an example, a finding and extracting data problem can be considered as a game that uses some of its factors as players, and solve the problem by playing the game. Collecting and using of data is also one of the most important issues in the game theory. This is very useful especially in iterative games. By processing and analyzing the obtained results of the opposing player in previous games, useful information can be extracted to use in subsequent decisions in iterative games. In fact, data mining can be used as a key technique for creation of a valid knowledge base for players of a game.

In this paper, a new framework is proposed to solve the materialized view selection problem using game theory. Cost of query processing and cost of view maintenance will be facing each other as two players of this game. Players' strategy can be one of the proposed methods to solve the materialized view selection problem. In this framework, the game continues until equilibrium is reached. After reaching equilibrium, the set of appropriate views will be specified directly or indirectly. Based on the proposed framework, a new Game Theory based Materialized View selection method (GTMV) is presented. The proposed method is a greedy algorithm that exploits MVPP structure to evaluate and select appropriate views. GTMV method uses static strategy for view selection. Since no limitation is considered in the problem, static view selection is very appropriate. It should be noted that the framework and the proposed method are flexible and expandable, and with some modification, the proposed method can be used to solve the dynamic view selection problem. Using game theory, the proposed method achieves better results in terms of time.

The rest of paper is organized as follows. In Section 2 the most important materialized view selection strategies are reviewed. Some preliminaries and problem definition is given in Section 3. The proposed framework and the new game theory based materialized view selection method is raised in Section 4 and is explained using an example. The proposed method is compared with other existing methods and experimental results are shown and evaluated in Section 5. Finally, the work is concluded in Section 6.

2. Related works

In this section, we first represent important existing works which are related to materialize view selection. Then, we explain important game theory applications in data science.

2.1. Materialized view selection strategies

Many heuristics for solving materialized view selection problem have been represented in the literature. Some of these methods are based on deterministic algorithms and produce optimized answers. The others methods have exploited soft computing techniques to find semi-optimal answers in shorter executing time compared to deterministic strategies. There exists a third category of materialized view selection methods which combine algorithms and techniques of two mentioned categories to solve the problem in reasonable time with high accuracy. Several deterministic algorithms to solve the problem of materialized view selection have been presented so far. Many of them use greedy strategy to solve the problem. The greedy algorithm of Yang et al. (1997) has used MVPP structure to select optimal set of materialized views. The algorithm has minimized both cost of query processing and cost of view maintenance, and has not considered any limitation. The proposed method of Roy et al. (2000) selects the appropriate set of views using a greedy heuristic and AND–OR DAG structure. This method only minimizes the cost of query processing and does not care to cost of view maintenance. An extension of this method has been represented in Mistry et al. (2001), in which, view maintenance cost is also considered by the extended method. No limitation is also considered in this extended method. A framework and a greedy algorithm for view selection for materialization using an AND–OR DAG and considering storage space limitation is proposed in Gupta (1997). The method of Harinarayan et al. (1996) has selected desired views from a data cube lattice considering memory space restriction. View maintenance cost has been ignored and only query processing cost has been minimized in this method. Both of query processing and view maintenance costs have been minimized by query rewriting based algorithm of Theodoratos and Sellis (1997), in which, no limitation has been considered. Another query rewriting based method for view selection has been represented in Jogekar and Mohd (2013), which has considered storage space limitation. In Wu et al. (2013), an efficient method for selecting materialized views from workload of XPath has been proposed, and in Katsifodimos et al. (2012), an efficient method for selecting materialized views from XQuery's workload of XML data has been introduced. Bag based and bag-set semantic based materialized view selection problems have been introduced in Afrati et al. (2014), and query rewriting based algorithm has been represented to solve