IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-9, Sept. 2018 E-ISSN: 2347-2693

A Hybrid Approach To Solving The View Selection Problem In Data Warehouse

Mohammed El Alaoui[1*], Karim El Moutaouakil[2], Mohamed Ettaouil [3]

[1] Modelling and Scientific Computing Laboratory, University Sidi Mohammed Ben Abdellah, Fez, Morocco
[2] Modelling and Scientific Computing Laboratory, University Sidi Mohammed Ben Abdellah, Fez, Morocco
[3] National school of applied sciences Al-Hoceima (ENSAH) BP 03, Al-Hoceima, Morocco

[*]*Corresponding Author: md.elalaoui@gmail.com*

Available online at: www.ijcseonline.org

Accepted: 20/Sept/2018, Published: 30/Sept/2018

***Abstract*—** A data warehouse is a centralized repository of information from one or more data sources. The amount of big data that arrives in data warehouse typically comes from transactional systems and other relational databases. Often the data is stored in the form of materialized views in order to improve the performance of query execution in data warehouse. One of the most important techniques for improving query optimization performance is the selection of views to materialize. In this paper, the views selection problem is modelled as constraint satisfaction and optimization problem. The exact method standard may take a considerable amount of time in order to find an optimal solution. To address this limitation of the exact method, we proposed an approach based on consistency techniques and systematic search techniques to select an optimal set of views for materialization. This proposed approach improves the quality of execution time for selecting an optimal set of views to materialize.

***Keywords*—** Data warehouse, view selection problem, constraint satisfaction and optimization problem, hybrid approach, exact method

I. INTRODUCTION

Data growth worldwide has pushed companies to use the data warehouse (DW) as a strategic technology for decision-making and market research. Data warehouse is a subject-oriented, integrated, non-volatile and time-variant collection of data that supports management's decision of any given entity. To increase the efficiency of the queries used in the DW, and to avoid the direct and continuous access to the source data, we must adopt the technique of the materialization. This technique can be seen as an intermediary that can respond to any request concerned by this materialization.

Materialization is a powerful approach for optimizing query execution time. However, the materialization of all or none of the views can generate two opposite cases. Better query performance can be achieved by total materialization, but the cost of maintenance is higher. In the other case, the non-materialization of the views may be better at the maintenance level but with a very high processing cost. To propose an optimal solution, we must then be aware that neither the space where the views can materialize nor the time of their maintenance are unlimited, we are confronted with another problem that of the choice on which it is better materialized.

The view selection problem is one of the most discussed topics in the literature for the optimal choice of materialized views in DW to improve query performance. That is in fact NP-hard problem because of the fact that the solution space grows exponentially as the problem size increases [1,2].

In the literature, four representations have been used as search space for the problem of selecting views in DW. Multidimensional Lattice cube representations of views [3,4,5,6,7].

AND-OR view graph representations of views [8,9,10]. Multiple view processing plan representation [11,12] and data mining representation [13,14]. The Multidimensional Lattice cube has been used to express the dependencies between different cells or views of the data cube. Graphically, this multidimensional lattice cube representation is composed of a set of nodes that are the views, and the arcs that represent the dependency between the views. Anjana Gosain et al. proposed particle Swarm optimization algorithm for materialized cube selection [15]. AND-OR view graphs representations were introduced to represent all the possible ways to generate warehouse views such that the best query path can be utilized to optimize query. Imene Mami et al. have chosen this framework for the selection of materialized views [8]. Multiple views processing plan is

constructed using all common or similar subexpressions among the queries for the view selection problem. Roozbeh Derakhshan et al. chose to use the simulated annealing algorithm to improve query performance by selecting views in a data warehouse [12]. Data mining representation, this approach is based on detection of common sub-expressions, and represented workload as a binary matrix, in this matrix, each row represents a query and each column is an attribute. This data mining techniques is used for view selection problem. Kamel Aouiche et al. have used data mining techniques and applied it to this matrix to obtain a set of candidate views for materializing [13].

In this paper, we solve the view selection problem using hybrid approach. Firstly, optimal MVPP is proposed as a search space. Secondly, the view selection problem is modelled in term of an original constraint satisfaction and optimization problems. Thirdly, hybrid approach that combines between consistency technique and systematic search technique to solve the proposed model.

The rest of the paper is organized as follows: In the second section, we propose a mathematical modeling for the view selection problem. In the third section, we describe hybrid approach. In the fourth section, describes our experimental results. The fifth section, concludes the paper.

II. MATHEMATICAL MODELING TO SOLVING THE VIEW SELECTION PROBLEM

Constraints satisfaction problem (CSP) provides a formalism for many real problems. The resolution of CSP consists of finding an assignment of values to variables, subject to a set of constraints. Sometimes, the ideal solution is not only to meet all the constraints but also to choose the best one. This justified the extension of CSP to constraint satisfaction and optimization problems CSOP via the introduction of a cost function [16]. The task we are interested in is to find a complete assignment satisfying all the constraints and minimizing the overall cost. CSOP introduces a cost function whose value depends on the values assigned to the variables. This CSOP has two objectives: Firstly, to face and satisfy all constraints by suggesting solutions, and secondly, to choose the most optimized solution amongst them. The CSOP can be solved by the exact algorithms such as generate and test, backtracking etc.[17]. But unfortunately these algorithms require a lot of time. In order to deal with this problem, we have proposed a hybrid approach which consists in using consistency technique and systematic research. In first time, we present a transformation of a view selection problem into constraint satisfaction and optimization problem (CSOP). In second time, a hybrid approach is proposed to solve the CSOP to ensure the optimal solution especially for small VSP. The VSP is modelled as follows:

CSOP is a quintuplet $X=(X,D,C,R,f)$ defined by:

$X=\{x_1,x_2,\cdots,x_n\}$ is a set of n variables (n views).

$D=\{D_1,D_2,\cdots,D_n\}$ is a set of n discreet and finite domains, where D_i is the set of values associated with variable x_i.

$C=\{C_1,C_2,\cdots,C_m\}$ is a set of m constraints: where any constraint C_i concerns a subset of variables.

$R=\{R_1,R_2,\cdots,R_m\}$ is a set of m relations, where each relation R_i is defined by a subset of the Cartesian $D_{i1}\times D_{i2}\times\cdots\times D_{ik}$ product corresponding to the set of possible value combinations for C_i, k being the number of variables involved in C_i.

f is a cost function defined by:

$$f=\sum_{q\in Q} Cost_q(v)+\sum_{m\in M} Cost_m(v)$$

Where

M is the set of materialized view (m),

Q is the set of query (q),

$Cost_q(v)$ is the cost of query processing,

$Cost_m(v)$ is the cost of maintenance.

III. MULTIPLE VIEW PROCESSING PLAN

In this work, we used multiple views processing plan as a search space to obtain an optimal set of views to materialize. Multiple views processing plan representation is used to exploit the common sub-expressions that can be detected among the queries. The leaf nodes correspond to the base relations and the root nodes corresponds to warehouse queries. The process of building the search space can be divided into two phases: The first phase is the identification of common tasks among a set of queries and prepares a small set of alternative plans for each query. The second phase generates a global execution plan that will produce the answers for queries as they execute. In Figure 1, the global MVPP is constructed by combining the local query plan the first query with second query. Each query has multiple execution plans. Each query has one or more plans. The search space of MVPP is obtained by selecting a plan for each request. Optimal MVPP is the one that has a lower total cost. In each graph, the query access frequencies are labelled on the top of each query node. And for each node except the root (query node) and leaf (base relation node) nodes, there are two data associated with it. The left stands for the query operator and the right stands for the cost to generate the nodes from base relations. The view selection problem has been modelled as constraint satisfaction and optimization problems. The model consists of variables, domains, constraints and the object function.

In this example, we represent the multiple view processing plan from two queries Q1 and Q2. Based on this obtained

plan, we apply the hybrid algorithm to select a set of views to materialize.

Q1. *select c_nation, s_nation, d_year,*
sum(lo_revenue)
from customer, lineorder, supplier, dates
where lo_custkey = c_customerkey
and lo_suppkey = s_suppkey
and lo_orderdatekey = d_datekey
and c_region = 'ASIA'
and s_region = 'ASIA'
and d_year >= 1992 and d_year <= 1997
group by c_nation, s_nation, d_year ;

Q2. *select c_city, s_city, d_year,*
sum(lo_revenue)
from customer, lineorder, supplier, dates
where lo_custkey = c_customerkey
and lo_suppkey = s_suppkey
and lo_orderdatekey = d_datekey
and c_nation = 'UNITED STATES'
and s_region = 'ASIA'
and d_year >= 1992 and d_year <= 1997
group by c_city, s_city, d_year ;

Based on MVPP presented in Figure 1, each node presents a variable associated with its domain of definition.
Variable $X = \{tmp1, tmp2, \cdots, tmp13\}$ and
Domain $X = \{D_1, D_2, \cdots, D_{13}\}$ such as $D_1 = D_2 = \cdots = D_{13} = \{0,1\}$
The constraint is defined as follows: a node (view) cannot be selected with its descendants at the same time.

A binary constraint network is a constraint satisfaction and optimization problem for which all constraints are of arity two (i.e. is a relation over exactly two variables) i.e. $C_{i,j} : x_i x_j = 0$

The binary relation is defined by:

$R_{1,4} = R_{2,3} = R_{3,4} = R_{9,10} = \{(0,0),(0,1),(1,0)\}$

$R_{7,8} = R_{8,12} = R_{6,11} = R_{11,13} = \{(0,0),(0,1),(1,0)\}$

A non-binary constraint is a constraint that involves more than two variables i.e. $C_{i,j,k} : x_i x_j x_k = 0$.

The objective function f is defined by a sum of two objective functions, the first function represents cost of query processing and the second represents cost of maintenance.

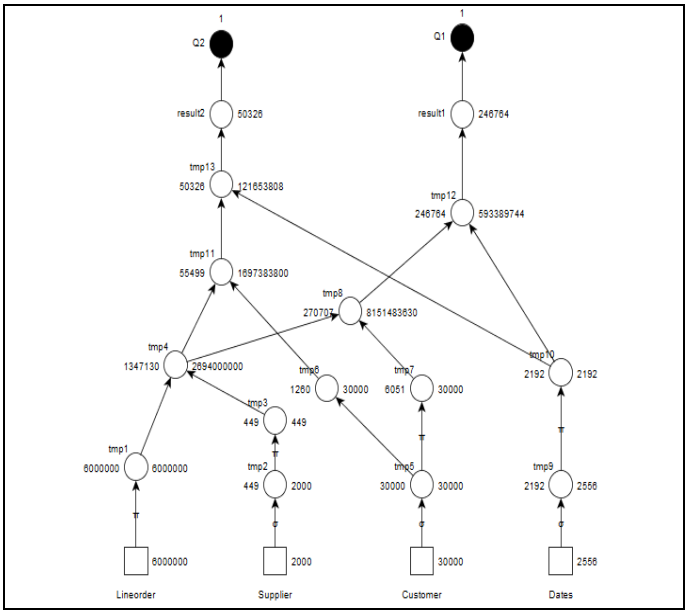

Figure 1. MVPP after all query Q1-Q2 are merged

In the data warehouse, selecting views using the MVPP search space relies on the optimal choice of views such that the total cost for query processing and view maintenance is minimal. The sum cost is calculated from the possible combination of nodes. The node in the search space is represented by a view. The cost metrics for selecting materialized views are based on the following costs: Cost of query processing is the frequency of the query multiplied by cost of query access from materialized views. Cost of view maintenance is equal to the cost of constructing the view in response to the changes in the base relation. Total cost is equal to the sum of the cost of query processing and the cost of view maintenance. The queries are represented by the root nodes, the base relationships are the leaf nodes, and the other intermediate nodes are selection, projection, join, and aggregate views that form a given query [18]. The sum cost is calculated from the possible combination of nodes. A search space of n nodes required 2^n combinations to find the optimal set of the views to materialize. Suppose that there are some intermediate nodes to be materialized. For each query, the cost of query processing is query frequency multiplied by the cost of query access from the materialized node(s). In Figure 1, if node tmp4 is chosen to be materialized, the query processing cost for Q1 is 1*(246764+ 593389744+ 2192+ 2556+ 8151483630+ 30000+ 30000+ 1347130). The maintenance cost for the materialized view is the cost for the process of updating a materialized view in response to the change in the base relations. The view maintenance cost of temp4 is 2*(2694000000+ 449+ 2000+ 6000000). The purpose of the selection of views is to improve the performance of the requests, by materialization in the data warehouse.

IV. HYBRID APPROACH

The proposed hybrid approach is based on a systematic search that attempts to find a solution by systematically searching the search space and on the consistency technique that is used as a pre-processing step where inconsistencies are detected and eliminated, before starting the search or during the search process itself, to reduce the nodes to instantiate in the tree.

A. Systematic search

Systematic search is to go through the search space until a solution is found or prove that there is no solution. The possible combinations of assigning values to variables in a CSP give rise to a search space that can be represented as a search tree or graph. Each node in the search tree represents a partial assignment of values to a set of variables. There are several systematic search algorithm like generate & test, backtracking etc. [19].

B. Consistency techniques

Consistency techniques introduced for the first time in artificial intelligence to improve the efficiency of image recognition programs [20]. In the literature, several consistency techniques have been proposed as ways to improve the efficiency of search algorithms. These techniques are used as pre-processing steps where inconsistencies are detected and eliminated, before starting the search or during the search process itself, in order to reduce the nodes to be instantiated in the search tree. We can differentiate between different levels of consistency such as node consistency, arc consistency or path consistency [21]. Algorithms that achieve such levels of consistency eliminate the instantiations of values in the domains of incompatible variables, that is, they remove the nodes of the search tree, which cannot participate in any solution. By applying consistency algorithms, we do not guarantee that all remaining variable-value pairs are part of a solution; practice has shown that they can be very useful as a pre-processing step, to reduce complexity of CSPs and also during research, to reduce the search space.

C. Consistency techniques

The generate and test algorithm is the most commonly used technique because it instantiates each of the possible values on the variables and systematically traverses the entire search tree [22]. Unfortunately, this method has the disadvantage of being very slow when searching for a solution, because it generates assignments that do not respect the constraints, which results in a loss of time and cost. To remedy the problem of this method, we have proposed a hybrid approach that seeks to detect future inconsistencies even earlier.

Algorithm 1 describes the process of hybrid approach applied to the resolution of CSOP, where $V[n]$ represents the vector of assignments to the variables $(x_1,x_2,\cdots,x_n)$ of the views selection problem.

Algorithm 1: Hybrid_algorithm

Initialization: Hybrid_algorithm $(1;V[n])$

Début

Procedure: Hybrid_algorithm $(k;V[n])$

$V[k]$ = Selection (d_k) ; Select a value from domain of d_k to be attributed to variable x_k

If Verification $(k;V[n])$ then

  If $k = n$ then

    Return $[n]$; It's a solution

  Else

    Hybrid_algorithm $(k+1;V[n])$

  End If

Else

  If stay_value (d_k) then

    Hybrid_algorithm $(k;V[n])$

  Else

    If $k = 1$ then

      Return false;

    Else

    Hybrid_algorithm $(k-1;V[n])$

    End If

  End If

End If

End Hybrid_algorithm

V. SIMULATION RESULTS

The experiments are conducted on SSB benchmark. This benchmark is derived from the TPC-H with scale factors of 1GB. TPC-H is the benchmark of the Transaction Processing Performance Council (TPC) for decision support. The SSB benchmark contains of one large fact table LINEORDER and four dimensions tables CUSTOMER, SUPPLIER, PART and DATE [23]. In order to determine a suitable set of views that minimizes the total cost associated with the materialized views, in conjunction with MVPP framework, in this sense, a hybrid approach is applied to solve the view selection problem. The defined model has been implemented in the visual studio solver to validate the expected results presented in this article.

Table 1. MVPP, cost of query processing, cost of maintenance and total cost

	Cost of query processing	Cost of maintenance	Total cost
Optimal set materialized views (13views)	15964141229	7197	15964148426
Optimal set materialized views (16views)	12387214367	5400069646	17787284013
Optimal set materialized views (20views)	1557801686572	3400000	1557805086572
Optimal set materialized views (22views)	1546558528754	182773100280	1729331629034

Table 2 Comparison between hybrid approach and generate and test

	Generate and test method		Hybrid approach	
	iteration	time	iteration	time
set materialized views (13views)	8192	3,6309224s	858	0,5953907s
set materialized views (16views)	65536	36,5020063s	3126	1,839106s
set materialized views (20views)	1048576	701,6086125s	28776	11,2178131s
set materialized views (22views)	4194304	2406,6279004s	88112	37,747951s

In order to validate the proposed approach, some experiments are effectuated to solve some typical problems of the materialization of the views. These experiments are effectuated in personal computer with a 2GHz processor and 2GB RAM. This approach is implemented by visual studio language. The performance has been measured in terms the minimum obtained cost. In comparison with generate & test algorithm, the optimum cost obtained by our hybrid approach is very interesting. Moreover, these results are obtained in the minimum time (See Table 2). For instance, when solving problem instance of set materialized 22 views, our approach required only 37.74 seconds whereas generate and test algorithm is executed in 2406,62 seconds. Therefore, the Table 1, show the optimal total cost of materialized views obtained by using hybrid approach and generate & test algorithm. The optimal total cost of the materialized views refers to the sum of total query processing & maintenance cost of views.

VI. CONCLUSION

The view selection problem is considered one of the essential elements in the design of a data warehouse. The goal of this problem is to minimize the total cost which is the sum of the costs of query processing and the maintenance costs of the views. In this article, we proposed a hybrid approach for selecting an optimal set of views to materialize in the data warehouse. This hybrid approach is based on a combination of systematic search techniques and consistency techniques in order to predict the violation of a constraint prior to instantiation. In this context, hybrid approach is proposed on the basis of this notion of verification. First, we proposed a problem selection model for constraint optimization and constraint optimization. Then, we solved this model by a hybrid approach. The overall results demonstrated the effectiveness of the proposed algorithm on a naive approach with an experimental performance study. For future work, we are going to study the multi-objective optimization problem for the views selection problem and also many improvements could be made, especially regarding the order of variables and values.

REFERENCES

[1] H. Gupta and I.S. Mumick, "*Selection of Views to Materialize Under a Maintenance Cost Constraint*", Proc. 7th Int. Conf. Database Theory, vol. 13, pp. 453–470, 1999.

[2] H. Gupta and I.S. Mumick, "*Selection of views to materialize in a data warehouse*", IEEE Trans. Knowl. Data Eng., vol. 17, no. 1, pp. 24–43, 2005.

[3] D. Yang, M. Huang, and M. Hung, "*Efficient Utilization of Materialized Views in a Data Warehouse*", PAKDD 2002 Adv. Knowl. Discov. Data Min., pp. 393–404, 2002.

[4] G. Gou, J.X. Yu, and H. Lu, "*A* search: An efficient and flexible approach to materialized view selection*", IEEE Trans. Syst. Man Cybern. Part C Appl. Rev., vol. 36, no. 3, pp. 411–425, 2006.

[5] T.V.V. Kumar and S. Kumar, "*Materialized View Selection Using Simulated Annealing*", Int. Conf. Big Data Anal., pp. 168–179, 2012.

[6] C.S. Park, M.H. Kim, and Y.J. Lee, "*Finding an efficient rewriting of OLAP queries using materialized views in data warehouses*", Decis. Support Syst., vol. 32, no. 4, pp. 379–399, 2002.

[7] J. Chang and S. Lee, "*Extended conditions for answering an aggregate query using materialized views*", Inf. Process. Lett., vol. 72, pp. 205–212, 1999.

[8] I. Mami, R. Coletta, and Z. Bellahsene, "*Modeling view selection as a constraint satisfaction problem*", Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 6861 LNCS, no. PART 2, pp. 396–410, 2011.

[9] D. Theodoratos, "*Detecting redundant materialized views in data warehouse evolution*", Inf. Syst., vol. 26, no. 5, pp. 363–381, 2001.

[10] T.V.V. Kumar and S. Kumar, "*Materialised view selection using differential evolution*", Int. J. Innov. Comput. Appl., vol. 6, no. 2, pp. 102–113, 2014.

[11] M. El Alaoui, K. El moutaouakil, and M. Ettaouil, "*Weighted constraint satisfaction and genetic algorithm to solve the view selection problem*", International Journal of Database Management Systems (IJDMS), Vol.9, No.4, August 2017.

[12] R. Derakhshan and F. Dehne, "*Simulated Annealing for Materialized View Selection in Data Warehousing Environment*", 24th IASTED Int. Conf. Database Appl., pp. 89–94, 2006.

[13] K. Aouiche and J. Darmont, "*Data mining-based materialized view and index selection in data warehouses*", J. Intell. Inf. Syst., vol. 33, no. 1, pp. 65–93, 2009.

[14] K. Aouiche, P.-E. Jouve, and J. Darmont, "*Clustering-Based Materialized View Selection in Data Warehouses*", Lect. Notes Comput. Sci. Incl. Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinforma., vol. 4152 LNCS, no. 1, pp. 81–95, 2007.

[15] A. Gosain and Heena, "*Materialized Cube Selection Using Particle Swarm Optimization Algorithm*" Procedia Comput. Sci., vol. 79, pp. 2–7, 2016.

[16] M. Ettaouil, "*A 0-1 Quadratic Knapsack Problem for Modelizing and Solving the Constraint Satisfaction Problems*", Prog. Artif. Intell., vol. 1323, pp. 61–72, 1997.

[17] E.C. Freuder, "*A Sufficient Condition for Backtrack-Free Search*", J. ACM, vol. 29, no. 1, pp. 24–32, 1982.

[18] S. Chakraborty, J. Doshi,"Deriving Aggregate Results with Incremental Data using Materialized Queries",International Journal of Computer Sciences and Engineering,Vol.-6, Issue-5, May 2018

[19] R. Barták, M.A. Salido, and F. Rossi, "*Constraint satisfaction techniques in planning and scheduling*", J. Intell. Manuf., vol. 21, no. 1, pp. 5–15, 2010.

[20] K.S. Joo, T. Bose, and G.F. Xu, "*Image Restoration Using a Conjugate Gradient-Based Adaptive Filtering Algorithm **", vol. 16, no. 2, pp. 197–206, 1997.

[21] O. Lhomme, "*Consistency techniques for numeric CSPs*", Ijcai, pp. 232–238, 1993.

[22] F. Manya and C. Gomes, "*Solution Techniques for Constraint Satisfaction Problems*", Intel. Artif., vol. 7, no. 19, pp. 243–267, 2003.

[23] P.O. Neil, B.O. Neil, and X. Chen, "*Star Schema Benchmark - Revision 3*", Tech. rep., 2009.

Authors Profile

Mohammed El Alaoui is a PhD student in the Laboratory of Modeling and Scientific computing at the Faculty of Sciences and Technology of Fez, Morocco, he is a member of Operational Research and Artificial Intelligence. He works on Neural Network, constraint satisfaction problem, Query Optimization in data Warehouse, and relational database.

Karim Elmoutaoukil is a PhD in Artificial Intelligence from the University of Sidi Mohammed Ben Abdellah,Fez,Morocco. His main research topics are neural networks, optimization, clustering and machine learning. He has published an big part in different congress and journals. He is a member of Scientific Committee for different international congress. He is a Professor at National School of Applied Sciences of Al Hoceima, University Mohammed First, Box 03, Al Hoceima, MOROCCO.

Mohamed Ettaouil is a Professor at Faculty of Science and Technology of Fez, University Sidi Mohammed ben Abdellah, Fez, Morocco, is a member of Modeling and Scientific Computing Laboratory. His main research topics are neural networks, optimization, modelling and machine learning. He has published an big part in different congress and papers. He is a responsible of teams "Digital and Computer Engineering, Artificial Neural Networks and Learning (N2I-RNA)" at FST Fez. He is a member of Scientific Committee for different international congress.

