Abstract

Materialized views in Data Warehouse is a promising solution to speed up the analytical processing of huge volume of historical data for running decision support applications to detect business trends by mining the data. By materializing or storing information organized as a set of views within a data warehouse from different production databases avoid the accessing of the original data sources and thereby it increases the efficiency of the warehousing queries. The large number of computation on huge volume of data, and space required for materialization of the views make it impractical to materialize all possible views. Therefore, there is a need for selecting an appropriate set of views to materialize. This selection of views is commonly referred to as the view selection for materialization problem. Any change in the source data is to be reflected in the materialized views and hence the materialized views are to be maintained by synchronizing with the source data. Hence the trade-offs between query efficiency, materialized view maintenance cost, number of materialized views and/or their sizes make the view selection problem as one of the most challenging problems in data warehousing. The classic approaches presented in literature for handling this problem was deterministic algorithms with heuristic greedy approaches. With increase in dimensions of the data warehouses, the solution space increases exponentially and deterministic algorithm becomes un-scalable. The problem is NP-hard. Therefore, stochastic and evolutionary algorithms are found to be most suitable for selecting a set of views for materializing with optimum associated costs and benefits. In most of the recent works on this problem Genetic Algorithm (GA) and other randomized algorithms like Simulated Annealing (SA) have been applied for optimizing the costs of materialized views. Unfortunately, these approaches could not handle the issues and challenges related to this problem beyond a certain limit. The majority of the approaches consider the problem as mere single objective optimization problem to optimize summed up cost of all associated costs of materialized views in relational model based data warehouses ignoring trade-offs between different costs. Also they do not consider using an adaptive system for selecting views for materializing in case of data warehousing with very large semi-structured databases which support collective data

types like in case of Big data framework based data warehousing and data processing paradigm using Distributed File System at cluster of low cost commodity hardware.

We begin by defining the materialized view selection problem as a multi-objective optimization problem for optimizing query processing cost and materialized view maintenance cost with constraint on availability of space for materializing in conventional RDBMS based data warehousing. It addresses the issue of considering trade-offs between inter related costs that are to be optimized. An already established input method using multiple view processing plan based representation of the problem has been used for implementation. Though the classic Multi-Objective Genetic Algorithm has been already applied in this problem successfully, it has been observed that Differential Evolution (DE) algorithm outperforms Genetic Algorithms on many numerical single objective optimization problem. Therefore a version of the multi-objective Differential Evolution algorithm has been designed for adapting with binary encoded solution population for implementing in view selection problem. It also addresses the problem of loss of significant solutions while filtering out representative solutions out of large number of non dominating solutions of the multi-objective optimization problem.

With the advent of Big data and MapReduce programming paradigm, next we investigate on view selection problem for materializing in Big data framework. The problem has been defined as a multi-objective optimization problem for minimizing (1) the total query processing MapReduce cost, (2) materialized view maintenance MapReduce cost and (3) the number of views selected for materializing with constraint on minimum size of materialized views. The *Forma analysis* based multi-objective DE for binary encoded data, that has been already used in materialized view selection for conventional data warehousing, termed as MODE-BE, is modified and applied in designing a view selection and recommendation system for materializing in Distributed File System data warehouse framework by promoting diversity of solutions in solution vector space. The popular elitist multi-objective GA termed as NSGA-II is also applied on this problem to analyze performances between NSGA-II based systems and MODE-BE based recommendation system in view selection for materializing in Big data management framework.

Finally, For comparative analysis of performances of Multi-Objective Evolutionary Algorithms (MOEAs) with Simulated Annealing (SA) based techniques in materialized view selection, the basic Archived Multi-objective Simulated Annealing (AMOSA) algorithm is customized for applying in materialized view selection in MapReduce based distributed file system framework. So far available data warehousing technologies in Big data framework do not support materialized view. It is expected that this dissertation will be useful for future research in designing

and developing analytical processing applications for Big data warehouses.

Keywords: Data warehouse, View materialization, Materialized view selection, Query processing cost, Materialized view maintenance, Multi-objective optimization, Pareto front, Genetic Algorithm, Differential Evolution algorithm, Big data, MapReduce, Multi-objective Simulated Annealing (MOSA), Archived Multi-objective Simulated Annealing (AMOSA).