Query Frequency based View Selection

Mohammad Haider, Saudi Electronic University, Dammam, Saudi Arabia
T.V. Vijay Kumar, Jawaharlal Nehru University, New Delhi, India

ABSTRACT

View selection deals with the selection of appropriate sets of views capable of improving the response times for queries while conforming to space constraints. Materializing all views is infeasible, as the number of possible views is exponential with respect to the number of dimensions and, thus, would not fit within the available storage space. Further, optimal view selection is an NP-Complete problem. Thus, the only remaining alternative is to select a subset of views that reduce the query response time and fit within the available space for materialization. The most fundamental greedy view selection algorithm HRUA considers the size parameter for computing the Top-K views for materialization. In each iteration, it computes the benefit, with respect to size, of all non-selected views, and selects the one entailing the highest benefit for materialization. Though these selected views may be beneficial in respect of their size, they may not be capable of answering large numbers of future queries thereby becoming an unnecessary space overhead. Existing query frequency based view selection algorithms, which address this problem, have been compared in this paper. Experimental results show that each of these algorithms, in comparison to HRUA, are able to select fairly good quality views that provide answers to comparatively greater numbers of queries. Materializing these selected views would facilitate the business decision making process.

KEYWORDS

Analytical Queries, Decision Making, Greedy Algorithms, Materialized View Selection, Warehouse

INTRODUCTION

Globalization of businesses has led to voluminous amount of data being generated continuously over time. In this age of ever-changing data and a wants-driven economy, readily available and updated information plays a vital role in the formulation of optimal business strategies for gaining competitive advantage. To be, and remain, competitive in today’s volatile market, considerable efforts are required like conducting market research for identifying customer demands as against their needs. Exponential growth in the areas of information technology and information processing has been observed in the last few decades. Proper and timely availability of this processed information holds the key for businesses to survive. In order to meet this demand for information, the capture, and efficient storage, of the turbulent data that is to be processed for the purpose of analysis, should be the major focus. Such processed data generally proves useful for knowledge workers and/or decision makers in the decision making process. Availability of such data shall provide the business houses a substantive edge over their competitors.

With the advent of the era of technological enhancement in areas of software, analytics, hardware capabilities and data communication, most organizations have collected massive amounts of raw data. As a result, although most such organizations are data rich, they are lacking in cogent information (Gray & Watson, 1998; Han & Kamber, 2000) leading to valuable information getting lost inside humongous
data, resulting in organizations struggling for the appropriate information. Hence, the availability of the appropriate information to the appropriate individual gets delayed. This reinforces the need for the data available in the operational/other data sources to be converted into useful information to enable the knowledge workers, and/or the decision makers to access and extract the hidden essential patterns for making optimal decisions at the right time. In order to analyze such mammoth data, a sound decision support system is required for suggesting solutions to business problems/queries that are complex and unstructured. In this era of mounting competition amongst organizations, the need of the hour is to have systems capable of extracting, storing and analyzing the hidden information concealed in the mammoth data available. There are two ways to access data in data sources namely, the lazy or the on-demand approach and the eager or the in-advance approach (Widom, 1995). In the former, operational sources are accessed in response to queries thus reducing the storage overhead, whereas in the latter, data is pre-computed and stored beforehand resulting in reduced communication costs. Query response times for the on-demand approach are comparatively high as it explores the operational sources in response to the query posed. On the other hand, data is pre-computed and stored in the warehouse resulting in improved query response times in case of the eager or in-advance approach. Data warehouse is based on the eager or in-advance approach (Widom, 1995).

A data warehouse would enable organizations to survive in turbulent and highly competitive markets. Typically, in an organization, heterogeneous and independent information is spread across different geographical locations. For helping the process of decision making, a decision support system extracts, integrates and stores the relevant and required information from these data sources and stores it into a central repository, referred to as a data warehouse. In the year 1992 Ralph and Bill were working on a framework to figure out the way in which data could be collected from different business areas and meaningfully utilized at an enterprise level. Though working separately, both of them arrived at a similar set of principles for deploying data that is “Subject Oriented, Integrated, Non-volatile and Time-variant” (Inmon, 2003). Based on these principles, data warehousing extracts, integrates and stores the required and relevant information from the underlying data sources, into one central repository, referred to as a data warehouse. The purpose of a data warehouse is to support decision making (Inmon, 2003). The major challenge in storing pre-computed data in the warehouse is space constraint. This is referred to as the “curse of dimensionality” (Han & Kamber, 2000, Bauer & Lehner, 2003). As the number of views is exponential in the number of dimensions (Harinarayan et al., 1996), it becomes computationally infeasible to compute and store such a large number of views due to space constraints. A data warehouse stores historical information with the aim of providing answers for decision making queries, most of which are exploratory and analytical in nature and consume a lot of time for processing against the data warehouse resulting in a high query response time. The aim is to reduce this response time. One way to achieve it is by materializing views in a data warehouse (Roussopoulos, 1997; Zhou et al., 2008).

Unlike virtual views, which store only definitions, materialized views store data along with the definition. Materialized views are significantly smaller compared to a data warehouse (Roussopoulos, 1982). The purpose of a materialized view is to provide answers to most of the decision making queries posed on the data warehouse, in order to reduce the query response time. This necessitates that the materialized views contain relevant and required information for answering decision making queries. Major issues associated with materialized views are view selection (Boukra et al., 2007; Jiang et al., 2007; Lawrence and Rau-Chaplin, 2006; Qiu & Ling, 2000; Luo, 2007; Lawrence, 2006; Phuboon-ob & Auepanwiriyakul, 2007; Qiu & Ling, 2001; Bauer & Lehner, 2003; Ciferri & deSouza, 2000; Gofarelli & Rizzi, 2000; Roussopoulos, 1982; Theodoratos & Sellis, 1997; Tang et al., 2008; Yu et al., 2005; Yang et al., 1997; Horng et al., 1999; Mohania et al., 1998; Yin et al., 2007; Lin et al., 2007; Hanusse et al., 2009), view maintenance (Roussopoulos et al., 1997; Sarawagi, 1997;
Vintage Analytics and Data Warehouse Design
[www.igi-global.com/article/vintage-analytics-and-data-warehouse-design/120052?camid=4v1a](www.igi-global.com/article/vintage-analytics-and-data-warehouse-design/120052?camid=4v1a)