Abstract

Large Distributed Databases Systems have become quite common, but efficient query processing systems for them are still evolving. It is partially due to the NP Hard nature of the large scale distributed Query Processing Problem that involves many complex sub-problems, that themselves are NP-hard. Some of the sub tasks involved are: Data Fragmentation, Data Allocation, Operation Allocation, Operation Ordering, Load Balancing etc. Deciding a distributed database design and query processing strategies are the most challenging parts of this technology. It is a proven fact that finding an optimal execution strategy for a distributed query is computationally intractable. The research in this area dates back to more than three decades, but due to such a complex nature of the problem, the hunt for better solutions is still widely prevalent. Most of the research work involves looking for better mathematical programming models and finding better heuristics. All this is done in order to minimize the combined cost of storing the database, processing transactions against it and minimizing communication amongst network sites.

Hence query optimization’s first goal is to minimize data communication globally amongst various network sites and secondly minimize the disk accesses locally at a particular site. From global view point, Access Strategies in this thesis refer to the procedures applied on sub query allocation scheme to minimize the Communication Costs by minimizing the movement of data across the network sites. From a local viewpoint, it refers to Access Path Selection by minimizing the movement of data from secondary memory to main memory while performing local processing operations at a particular site.

Queries are categorized mainly in two groups, OLTP (Online Transaction Processing) and DSS (Decision Support System) queries. The former are generally repetitive in nature and hence optimizer should seek good thru-put by concentrating on reducing the total cost of the query. DSS queries more often seek response time optimization and hence use of parallel processing. There are two popular approaches to query optimization, first is reducing the Total Cost, second is reducing the Response Time of a distributed query. Total cost is the sum of all times incurred in processing the various operations of the query and inter-site communication. Response time is the time elapsed from origin to completion of the query.
In this thesis a simplified model for Total Cost Minimization for OLTP retrieval queries is proposed and implemented. A distributed database environment is simulated and query optimization is performed stochastically on Wisconsin benchmark queries using Genetic Algorithms. Experiments are performed by up scaling the size of the distribution and query complexity. This approach resulted in cost reduction of such OLTP retrieval queries by 30% as compared to previous composite models. Moreover the effect of types of genetic operators and genetic parameters value variations (Population size, Crossover% value, Mutation % value combinations etc) is studied and analyzed for distributed queries.