Statistical Analysis of DSS Query Optimizer for a Five Join DSS Query

Manik Sharma
Assistant Professor
Dept.of Computer Sc. &
App DAV University,
Jalandhar

Gurvinder Singh
Professor, DCS
Guru Nanak Dev University
Amritsar

Rajinder Singh
Professor and Head, DCS
Guru Nanak Dev University
Amritsar

Sarbjit Singh
System Manager
Guru Nanak Dev University
Amritsar

ABSTRACT

Statistics is a multidisciplinary subject which is used in different domains to mine decisive information from phenomenal volume of data. The aspiration of this paper is to examine the results of two different DSS query optimizers. The design of DSS query optimizer is based upon restricted exhaustive enumeration and a fusion of entropy and genetic algorithm called DSQ_REA and DSQ_ERGA. A five join DSS query is considered for assessing the results of DSQ_REA and DSQ_ERGA. The output of query optimizers is the plan that determine the location where the sub operations of the query would be executed. The objective of query optimizer is to select a query execution path that uses least amount of system resources. The outcomes of DSQ_REA and DSQ_ERGA are statistically analyzed by using different measures of descriptive statistics. Moreover, outlier analysis of the results has been carried out. In addition, the distribution of the results is also examined to reveal the deviation in the different query execution plans generated by DSQ_EA and DSQ_ERGA**.** With this statistical analysis, one is able to recognize the nature and distribution of different query execution paths generated with DSQ_REA and DSQ_ERGA.

Keywords

Statistical Analysis, DSS Query Optimizer, Entropy, Genetic Algorithm.

1. INTRODUCTION

Statistics is the branch of Mathematics which is used to unearth the trend or analyze the relationship between two or more variables. With the help of statistics, one is able to determine the characteristics of a variable or trend. In general, statistics methods are categorized as Inferential Statistics and Descriptive Statistics. Inferential Statistics are used to analyze immense amount of data consisting of population in million, billion or even more. With this amount of population, it is not realistic to get the data for each item of a population. The objective of inferential statistics is to infer about the nature of complete population using the partial data drawn from population. In simple words, an analyst will collect the sample from the population and deduce the nature of population from it. Estimation statistics and hypothesis testing are the common methods of inferential statistics. Descriptive Statistics represent the data in some meaningful manner. It analyzes the averages and dispersion of data. However, it does not attempt to describe the nature of population from where the sample has been taken. It attempts to analyze the distribution of data. Some of the common measures of descriptive statistics are measure of central tendency, dispersion, Skewness, Kurtosis and correlation [1][2][3].

Here, endeavor is to statistically analyze the results of two different DSS query optimization framework i.e. DSQ_REA and DSQ_ERGA. Before going in to details, let’s first understand what a DSS query is. A DSS query is a special type of distributed query which is fired to compute results that may assist in decision making. It plays substantial role in the organization and management of the data in the database system. DSS queries are normally focused to mine prodigious amount of data. The analysis of these types of queries is beneficial in developing a predictive model for the business organization. A DSS query can be classified as Reporting, Adhoc, Iterative OLAP and Data mining DSS query [4][5][12].

In common practice, DSS queries are used for read operations. These queries are used to provide aggregated or consolidated information. A DSS query is helpful in making strategic planning and future forecasting. Due to the decentralization of data and the complexity of query it becomes mandatory to optimize the DSS query in distributed database system. The distributed queries can be optimized by minimizing either the Total Costs or Response Time of a query. Total Costs are optimized for increasing the throughput of the system. On the other hand, Response Time is optimized to speed up the execution process of a query[6][7][8].

2. PROBLEM DEFINITION

The aim of this work is to statistically analyze the results of two different DSS query optimizer based upon exhaustive enumeration approach (REA) and a hybrid combination of entropy and restricted genetic approach (ERGA). The results are analyzed based upon a five join DSS query. Different descriptive statistical measures like mean, mode, variance, range, skewness and kurtosis are computed and compared. In addition, the endeavor is made to find outliers in the results of query execution plans generated by both DSQ_REA and DSQ_ERGA. Finally, the distribution of the results is examined to know the variation in the different possible combination of query execution plans.

3. STATISTICAL ANALYSIS OF FIVE JOIN DSS QUERY

A 'Five-Join DSS' query has been considered and optimized using REA and ERGA assuming Total Costs of the distributed DSS query. The composition of the 'Five Join DSS' query and corresponding query tree is presented below.

Statistics of Query

Number of Base Relations	**:**	**06**
Number of Sites	**:**	**10**
Number of Join Operations	**:**	**05**
Number of Selections	**:**	**06**
Number of Projections	**:**	**06**
Number of Operations	**:**	**17**
Number of Fragments	**:**	**22**

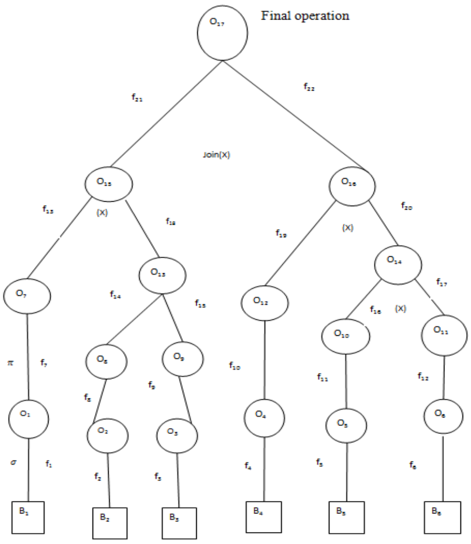

Figure 1: Five Join DSS Query

Both DSQ_REA and DSQ_ERGA generates number of query execution schemes. Different statistical measures viz. mean, mode, standard deviation, range, coefficient of variation etc. has been computed for the Total Costs of different query execution schemes generated by DSQ_REA and DSQ_ERGA.

Table 1 shows the values of distinct statistical measures, when the 'Five-Join DSS' query is optimized using DSQ_REA and DSQ_ERGA. From Table 1, it is perceived that distribution of the Total Costs of different query execution schemes as given by DSQ_REA and DSQ_ERGA is not normal as mean, median and mode value do not lie on a single point.

Table 1: Measures of Central Tendency and Dispersion

Measures	Total Costs of Five Join DSS Query using DSQ_REA	Total Costs of Five Join DSS Query using DSQ_ERGA
Mean	2563780.50	2570013.84
Median	2565320.00	2563115.00
Variance	3835897140.339	1030807733.46
Minimum	2422235	2459420
Maximum	2777250	2733560
Skewness	.230	2.818
Kurtosis	-.316	8.702
Optimal Value of Total Costs	2422235	2459420

Furthermore, the mean value of different query execution plans of a 'Five Join DSS' query as generated by DSQ_REA is more close to the optimal value of Total Costs as compared to the median of the query execution plans. The range of Total Costs of DSS query is 355015. From the values of minimum, maximum and range, it is obvious that best query execution scheme generated by REA is 87% more optimal than the worst one. Outlier is one of the important parameter of statistics that determines values which are far away from the central value. Here, outliers are those query execution schemes which should be avoided as they consume lots of system resources. Obviously, these values cannot be the candidate of solution. Table 2 (a), shows the outlier found in the various query execution plans generated by DSQ_REA.

Table 2 (a): Outliers in Total Costs of DSS Query using REA

Extreme Values				
			Case Number	Value
Total Costs using REA	**Highest**	1	536623	2777250
		2	556623	2775210
		3	216623	2774955
		4	496623	2774955
		5	376623	2774700

From Table 2 (a), it is clear that all the values of Total Costs greater than 2774700 are outliers. Here, outliers represent the worst scenario of the Total Costs of DSS query.

Table 2 (b) : Outliers in Total Costs of DSS Query using ERGA

Extreme Values				
			Case Number	Value
Total Costs using ERGA	**Highest**	1	78	2776490
		2	327	2779872
		3	729	2778952
		4	854	2779155
		5	989	2779845

Table 2(b) represents that all values of Total Costs obtained using ERGA greater than 2776490 are outliers. Furthermore, in case of REA, the value of β2(Kurotosis) is -0.316 which is less than zero. Therefore, the curve is platy kurtic and is more flat in nature. Figure 2 shows the frequency histogram of assorted values of Total Costs of ‘Five-Join DSS’ query obtained using DSQ_REA.

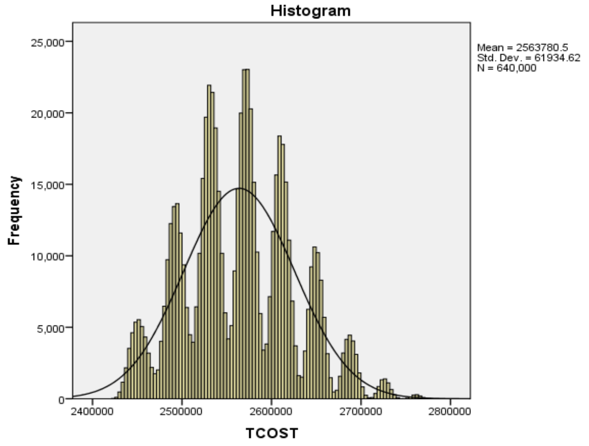

Figure2 : Histogram of various Query Execution Plans generated using REA

In case of DSQ_ERGA, the value of median is more close to the optimal value of Total Costs of a DSS query as compared to its mean value. The value of minimum, maximum and range of Total Costs for ‘Five-Join DSS’ query is 2422235, 27777250 and 350105 respectively. From the values of minimum, maximum and range, it is obvious that best query execution scheme generated by ERGA is 89% more optimal than the worst one.

Figure 3(a) and 3(b) demonstrates the comprehensive analysis of different descriptive statistical measures of the ‘Five Join DSS’ query obtained using DSQ_REA and DSQ_ERGA. From Figure 3(a) and 3(b), it is observed that mean, median and maximum values of Total Costs of DSQ_REA and DSQ_ERGA are almost same. However, a significant difference has been observed in skewnewss, kurtosis and variance of data generated using DSQ_REA and DSQ_ERGA.

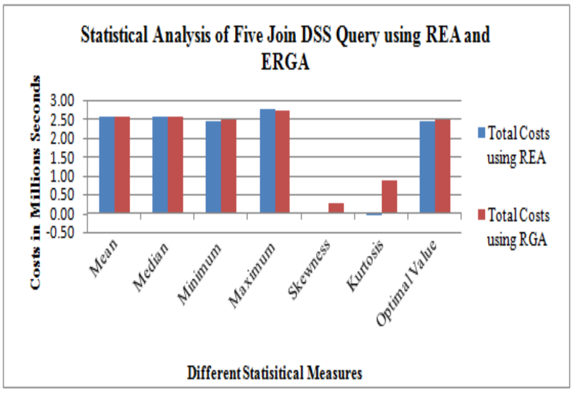

Figure 3 (a): Statistical Analysis of Five Join DSS Query

Figure 6.3 (b) illustrates the analysis of variance of ‘Five Join DSS’ query. The variance obtained in different query execution schemes of REA is very high as compared to the variance of ERGA. Thus, it is verified that the solution set of REA is spanned over a large range as compared to ERGA.

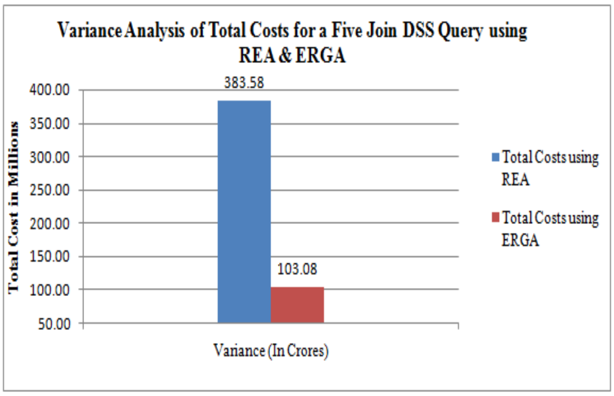

Figure 3 (b): Analysis of Variance of Five Join DSS Query

Unlike DSQ_REA, the value of β2 (Kurotosis) obtained using DSQ_ERGA, is 8.702 that is greater than zero. Therefore, the curve is leptokurtic and is more peaked in nature. The peakedness in the histogram of distinct values of Total Costs of the ‘Five-Join DSS query obtained using ERGA is witnessed.

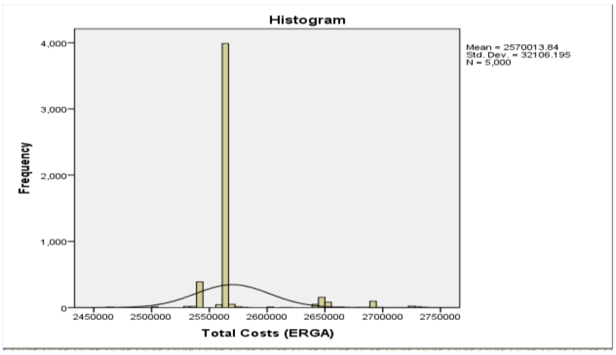

Figure 4 : Histogram of various Query Execution Plans generated using ERGA

From Figures 2, 3 and 4, it is observed that REA has lots of variations in the query execution plans as compared to the ERGA.

4. CONCLUSION

A number of common measures of descriptive statistics are computed and analyzed for two different DSS query optimizer framework viz. DSQ_REA and DSQ_ERGA. The best solution obtained using DSQ_REA and DSQ_ERGA is 87% and 89% better than the worst one. The value of Kurtosis of Total Costs of DSS queries obtained using REA is less than zero. Therefore, the Total Costs frequency histogram of different Operation Site Allocation Plans generated by DSQ_REA is Platy Kurtic in nature. However, the value of Kurtosis of Total Costs of DSS queries for different Operation Site Allocation Plans generated by DSQ_ERGA is greater than zero. Therefore, the Total Costs frequency histogram of different Operation Site Allocation Plans generated by DSQ_ERGA is Lepto Kurtic in nature. In common practice, the values of Mean and Median of Total Costs computed for DSQ_REA and DSQ_ERGA have immaterial variation. Significant Variation in Skewness and Kurtosis of Total Costs computed for REA and ERGA is observed.

5. REFERENCES

[1] http://sociology.about.com/od/Statistics/a/Descriptive-inferential-statistics.htm (As accessed on 15 Feb, 2016)

[2] SP Gupta. Statistical Methods. 43rd Edition 2014 Sultan Chand and Sons Publishers.

[3] Digambar Patri, D.N. Patri. Computer Mathematics and Statistical Methods. Second Revised Edition 2005, Kalyani Publishers.

[4] Manik Sharma, Gurvinder Singh, Rajinder Singh. “Design and Analysis of Stochastic DSS Query Optimizer in a Distributed Database System”. Egypitan Informatics Journal. doi:10.1016/j.eij.2015.10.003.

[5] Manik Sharma, Gurvinder Singh, Rajinder Singh and Gurdev Singh. 2015. “Analysis of DSS Queries using Entropy based Restricted Genetic Algorithm”. Applied Mathematics and Information Science. Vol. 9, Issue 5.

[6] Clark D. French. One Size Fits All- Database Architecture Do Not Work for DSS. ACM SIGMOD Newsletter1995:24-2:449-450.

[7] Narasimhaiah Gorla, Suk-Kyu Song. Subquery Allocation in Distributed Database using GA. Journal of Computer Science and Technology. 2010; 10-1:31-37.

[8] AhmetCosar, Sevinc Endvic. An Evolutionary Genetic Algorithm for Optimization of Distributed Database Queries. The Computer Journal. 2011; 54: 717-725.

[9] Peter M.G., Hevner Alan N., Yao Bing S. Optimization Algorithms for Distributed Queries. IEEE Transaction on Software Engineering1983.;9-1:57-68.

[10] Sangkyu Rho, Salvatore T. March. Optimizing Distributed Join Queries: A GA Approach. Annals of Operation Research 1997; 71:199-228.

[11] Manik Sharma, Gurdev Singh. Analysis of Joins and Semi Joins in Centralized and Distributed Database Queries. IEEE Xplore 2012; 978-1-4673-2647-6:15-20.

[12] TPC Benchmark DS, Version 1.1.0, April 2002 http://www.tpc.org (Accessed on June 2013).