Finding Top-k Dominance on Incomplete Big Data Using MapReduce Framework

Abstract—Incomplete data is one major kind of multi-dimensional dataset that has random-distributed missing nodes in its dimensions. It is very difficult to retrieve information from this type of dataset when it becomes large. Finding top-k dominant values in this type of dataset is a challenging procedure. Some algorithms are present to enhance this process, but most are efficient only when dealing with small incomplete data. One of the algorithms that make the application of TKD query possible is the Bitmap Index Guided (BIG) algorithm. This algorithm greatly improves the performance for incomplete data, but it is not designed to find top-k dominant values in incomplete big data. Several other algorithms have been proposed to find the TKD query, such as Skyband Based and Upper Bound Based algorithms, but their performance is also questionable. Algorithms developed previously were among the first attempts to apply TKD query on incomplete data; however, these algorithms suffered from weak performance. This paper proposes MapReduced Enhanced Bitmap Index Guided Algorithm (MRBIG) for dealing with the aforementioned issues. MRBIG uses the MapReduce framework to enhance the performance of applying top-k dominance queries on large incomplete datasets. The proposed approach uses the MapReduce parallel computing approach involving multiple computing nodes. The framework separates the tasks between several computing nodes to independently and simultaneously work to find the result. This method has achieved up to two times faster processing time in finding the TKD query result when compared to previously proposed algorithms.

Conclusion

In this paper, we proposed an algorithm to apply top-k dominating queries using MapReduce framework on incomplete big data. MapReduced Enhanced Bitmap
Indexed Guided algorithm (MRBIG) is the basis of the work that develops a new way to handle large incomplete data and uses the MapReduce framework to enable parallel computing to manage the problem faster. Throughout the paper, the single machine algorithm has been detailed, compared, and contrasted with the MRBIG algorithm. Based on the experiments, the single machine algorithm cannot be an optimal way for applying TKD queries on big files. Not being resource-efficient, process failure due to resource insufficiency, and having exponential processing time are among the major defects when it comes to finding top-k dominant values in massive incomplete data.

**SYSTEM REQUIREMENTS:**

**HARDWARE REQUIREMENTS:**

- System: Pentium IV 2.4 GHz.
- Hard Disk: 40 GB.
- Floppy Drive: 1.44 Mb.
- Monitor: 15 VGA Colour.
- Mouse: Logitech.
- Ram: 512 Mb

**SOFTWARE REQUIREMENTS:**

- Operating system: Windows 7/UBUNTU.
- Coding Language: Java 1.7, Hadoop 0.8.1
- IDE: Eclipse
- Database: MYSQL
REFERENCES

