Example of Project Proposal

Indexing techniques for big data: taxonomy and performance evaluation

FirstName Surname†  
Institution/University Name  
 email@email.com

ABSTRACT

The explosive growth in volume, velocity, and diversity of data produced by technology, such as sensors, mobiles, etc., and cloud applications has contributed to the abundance of data or ‘big data.’ Available solutions for efficient data storage and management cannot fulfill the needs of such heterogeneous data where the amount of data is continuously increasing. The objective of this paper is to investigate and examine the existing indexing techniques for big data. Taxonomy of indexing techniques is developed to provide insight to enable researchers understand and select a technique as a basis to design an indexing mechanism with reduced time and space consumption for BD-MCC. In this study, 10 indexing techniques is studied and compared based on some articles related to the topic. The indexing techniques’ performance is analyzed based on their characteristics and big data indexing requirements.

1 Introduction

For efficient retrieval and management, existing indexing solutions become inefficient with the rapidly growing index size and seek time and an optimized index scheme is required for big data. Regarding real-world applications, the indexing issue with big data in cloud computing is widespread in healthcare, enterprises, scientific experiments, and social networks. To date, diverse soft computing, machine learning, and other techniques in terms of artificial intelligence have been utilized to satisfy the indexing requirements, yet in the literature, there is no reported state-of-the-art survey investigating the performance and consequences of techniques for solving indexing in big data issues as they enter cloud computing. In taxonomy, indexing techniques are categorized based on three methods which are non-artificial intelligent (NAI), artificial intelligent (AI), and collaborative artificial intelligent (CAI) techniques. Traditional or non-artificial intelligence indexing approaches, such as bitmap indexes [1], graph query processing [2], and tree-based indexing, perform reasonably regarding volume, velocity, variety, variability, value, and complexity, but they fail to detect ‘unknown’ behavior or unknown big data.

2 Related work

State-of-the-art indexing techniques utilize large datasets with a variety of data types. They are designed to cope with certain data requirements; thus, they are suitable for specific situations. A hash-based indexing technique presented by Zhu et al. [3] is designed to be search-efficient in the context of high-dimensional data. The compact Steiner tree (CST), an extension of the Steiner tree, is developed to allow for improved keyword searches on relational databases [4]. This study discusses techniques that are currently available for data indexing and examines their applicability in data domains besides their suitability in some conditions like search cost, resource consumption, result accuracy, and type of queries to be applied.

**3 Research Questions**

Consequently, the intent of this project is to address the following:

• How to classify indexing techniques used in big data?

• What impact do the taxonomy design and the indexing techniques have based on big data indexing requirements?

**4 Approach**

The proposed approach is to take into account multiple platforms which are able to analyze unstructured data for the most efficient big data indexing techniques. Based on this review, hopefully analyses to what extent their pros and cons for a categorization of existing data management and indexing techniques for big data works, which helps to get more insights about CAI-based indexing methods. This approach is more prominent for data indexing and retrieval, since they are adaptable to large size data, one of the main issue for BD-MCC. CAI-based methods can provide satisfactory retrieval rate and accuracy of data retrieval in cloud, whereby data are continuously captured and utilized by end users.

REFERENCES

[1] Wu K, Shoshani A, Stockinger K (2010) Analyses of multi-level and multi-component compressed bitmap indexes. ACM Trans Database Syst 35(1):1–52. doi:10.1145/1670243.1670245.

[2] Cheng J, Ke Y, Fu AW-C, Yu JX (2011) Fast graph query processing with a low cost index. VLDB J 20(4):521–539.

[3] Zhu X, Huang Z, Cheng H, Cui J, Shen HT (2013) Sparse hashing for fast multimedia search. ACM Trans Inf Syst 31(2):1–24. doi:10.1145/2457465.2457469.

[4] Li G, Feng J, Zhou X, Wang J (2011) Providing built-in keyword search capabilities in RDBMS. VLDB J 20(1):1–19.