Indexing strategies of MapReduce for Information Retrieval in Big Data

ABSTRACT

In Information Retrieval (IR), the efficient strategy of indexing large dataset and terabyte-scale data is still an issue because of information overload as the result of increasing the knowledge, increasing the number of different media, increasing the number of platforms, and increasing the interoperability of platforms. Across multiple processing machines, MapReduce has been suggested as a suitable platform that use for distributing the intensive data operations. In this project, sensei and Per-posting list indexing (Terrier) will be analyze as they are the two efficient MapReduce indexing strategies. The two indexing will be implemented in an existing framework of IR, and an experiment will be performed by using the Hadoop for MapReducing with the same large dataset. In particular, this paper will study the effectiveness of two indexing strategies (Sensei & Terrier), and try to find and verify the better efficient strategy between them. The experiment will measure the performance of retrieving when the size and processing power enlarge. The experiment examines how the indexing strategies scaled and work with large size of dataset and distributed number of machines. The throughput will be measured by using MB/S (Megabyte per Second), and the experiment results analyzing the performance and efficiency of indexing strategies between Sensei & Per-posting list indexing (Terrier).

Keyword: Hadoop; Indexing; MapReduce; Sensei; Terrier