Performance evaluation of distributed indexing using Solr and Terrier information retrievals

ABSTRACT

The continuous growing datasets and the emergence terabyte-scale data pose great challenges to Information Retrieval (IR) systems. Tremendously, a large amount of data from various aspects is collected every day making the amount of raw data extremely large. As a result, indexing a large volume of data is a time-consuming problem. Therefore, efficient indexing of large collections is getting more challenging. MapReduce is a programming model for the computing of large document collections by distributing data and processing tasks over multiple computing machines. In this study, Solr and Terrier distributed indexing will be evaluated as they are the most popular information retrieval frameworks among researchers and enterprises. To be more specific, this paper will compare and analyze the distributed indexing performance over MapReduce for the indexing strategies of Solr and Terrier using 1GB, 3GB, 6GB, and 9GB datasets. In the experiments, the indexing average time, speedup, and throughput are observed as the number of machines involved in the experiments increases for both indexing frameworks. The experimental results show that Terrier is more efficient with large datasets in the presence of processing resource scalability. On the other hand, Solr performed better with small datasets using limited computing resources.

Keyword: Information retrieval; Indexing; MapReduce; Hadoop; Solr; Terrier