ABSTRACT

The storage space being demanded by the information resources is rapidly increasing resulting in the generation of high volume of digitally stored data. The quantity and scope of data is primarily in the form of massive document corpus databases that create many challenges in retrieving the necessary information. Latent Semantic Indexing (LSI) is a well-known technique in the domain of information retrieval. It is not only based on the keyword matching, but also employs statistics and algebraic computations, such that the noises could be cleaned. The troubles of synonymy and polysemy in the conventional techniques can be conquered based on the analysis of the terms related with the documents. Still, it is evident that LSI has problem to scale owing to the computing complication.

This thesis offers a Web Conscious Distributed Indexing Algorithm MapReduce-LSI which can improve the scalability using Hadoop structure based on the MapReduce dispersed computing model. It also addresses the overhead caused by the clustering algorithm. The assessments point out that MapReduce-LSI can result in noteworthy improvement compared with the other strategies on processing large scale of documents. The benefit of Hadoop is that, it sustains diverse computing environments with unbalanced load between nodes. This thesis proposes a genetic algorithm based on MR-LSI load balancing algorithm for balancing load in static and dynamic environments.

The genetic algorithm employs latent chromosomes there by achieves better time interval for executing load balancing so as to reduce overhead by controlling number of iterations. The result reduces overhead to large extent and it advances the performance of a cluster depending on the heterogeneity ranks as well.