Parallel Information Retrieval on A PC-Cluster Using Vector Space Model

Review by: Tawa Khampachua *{Tawa.K@Student.chula.ac.th}*

Nowadays, the volume of electronic text has been growing up at an exponential rate, nearly double size every six months. They become more expensive to manage with a conventional information retrieval system. Searching and indexing costs grow with the size of the underlying document collection. It means larger document collections invariably result in longer response time. As more documents are added to the IR system, performance may deteriorate to the point where that system is no longer usable.

To keep up with future retrieval requirement, we must turn to alternative architectures and algorithms. In this thesis we explore parallel and distributed information retrieval techniques. The application of parallelism can greatly enhance our ability to scale up traditional information retrieval algorithms and to support larger document collections.

The design of our information retrieval algorithm is based on the notable vector space retrieval model and implemented on a PC-Beowulf cluster, a network of personal computer. Large text collections are partitioned into manageable size so that each computing node in the cluster can handle it efficiently.

From the experimental result we found that our parallel retrieval system can be scaled up well with large collections which will be very useful for future information retrieval application.