要 旨

ペアワイズ計算の MapReduce 実装に
おけるデータ配置の処理時間への影響

前田 秀樹

分散ファイルシステム上のファイルを用いて計算する際，データがネットワークを介して転送されると処理効率の低下を招く．また，類似度検索の需要が高まっていることから，本研究では MapReduce で実装されたペアワイズ計算におけるデータ配置の処理時間への影響に着目した．MapReduce で実装されたペアワイズ計算の処理時間がデータ配置によって影響が出ることを，データノードと TaskTracker が同一の場合と異なる場合とで調査を行う．その結果，ブロックサイズとファイルサイズ次第では実際に影響が出ることが確認された．

キーワード　ペアワイズ計算, 分散ファイルシステム, MapReduce

Abstract

Influence of Data Placement on Processing Time in MapReduce-based Pairwise Computation

Hideki MAEDA

When we perform computation with files on distributed file systems, transfer of data over the network may cause lower performance. Since the demand for the similarity search is increasing, in this study we focus on the pairwise computation implemented in MapReduce, and study the effect of data arrangement on the processing time. We conduct experiments in two cases: when the tasktracker is on the datanode, and when the tasktracker is out of the datanode. As a result, we confirmed the effect on performance depending on the file size and the block size.

key words pairwise computation, distributed file system, MapReduce