要 旨

MapReduce フレームワークを用いた
木構造処理の実現に関する研究

川村　高之

木構造は，探索アルゴリズムやファイルシステム等でよく用いられている．近年では，XML 文書や HTML 文書といった木構造を利用したデータの大規模化が進んでいる．これにより，一台の PC では処理が終わらない，データが扱えないといった問題が発生し，処理の高速化が課題となっている．そこで，並列化することが重要となる．しかし，並列計算プログラムの作成は，逐次プログラムに比べて難しい．特に，木構造に対しては，ノードの親子・兄弟関係にも気を配らなければならない．よって，容易に木構造に対する並列プログラミングを行う方法が必要である．本論文では，大規模クラスタ向け並列計算フレームワークである MapReduce フレームワークを用いて木構造処理を実現する．そして，高速化が可能であるのか，どの程度効果があるのかを検証した．実験は，実際に実装したプログラムを 8 台構成のクラスタを用いて実行して処理時間を計測した．結果は，PC 台数を増やすことで，処理を高速化する事ができた．しかし，6 台以降は並列化の効果が薄くなっていた．原因は，ネットワークのオーバヘッドの増加と Hadoop のパラメータ設定のチューニング不足である．

キーワード　MapReduce 木構造

Abstract

Realizing Tree-Structure Processing on MapReduce Framework

Kawamura Takayuki

Tree structure are often used in large-scale data management and search algorithm etc. In recent years, has gotten bigger is data using tree structure such as XML and HTML documents. Problem occurs such as single PC can not end the process and handle data. Processing speed has question. Therefore, Parallelization is important. However, the creation of parallel programs is more difficult compared to the sequential program. In particular, the tree structure, have to care brother and parent-child relationship of node. Therefore, need an easy way to parallel programming for the tree structure. In this paper, realization of processing tree structure using the MapReduce framework, which is a parallel computing framework for large-scale clusters. And to verify it is possible to speed up, and how effective. Experiment, processing time is performed using cluster of 8 nodes implemented program. The result is by increasing the number of PC, able to speed up the process. But, 6nodes later is thinner of parallel effect. Cause is the lack of tuning of Hadoop configuration parameters and an increase in network overhead.

key words MapReduce Tree structure