要 旨

効率の良い Hadoop MapReduce プログラム生成に関する研究

宮崎　玲奈

Google によって提案された MapReduce は，大規模データ処理に対するフレームワークのひとつであり，そのオープンソース実装である Hadoop MapReduce は広く用いられている．MapReduce プログラムをより開発しやすくするための言語処理系として，Sawzall，FlumeJava，Pig，Hive などが提案されている．しかし，それらは主に，MapReduce モデル上で比較的容易に実装できるような，MapReduce のプログラミングモデルに近いアプリケーションを対象とするものであった．この問題に対し，本研究では，MapReduce のプログラミングモデルを意識することなく記述された，1 次元配列を操作する逐次プログラムを入力として，Hadoop MapReduce プログラムを生成する新しい生成系を提示する．

本研究で対象とする Hadoop MapReduce では，多様なパラメータや機能を提供しているため，同じ処理を行うプログラムであってもフレームワーク上では動作の異なる複数種のプログラムが記述可能である．このため，本研究では，はじめに，一次元配列上のデータ処理，特に配列の先頭から累積計算を行うような処理に対する Hadoop MapReduce 上の実装を 5 種類示す．そして, それぞれの実装の性能評価や，Hadoop のパフォーマンスへ影響する要素について検証を行う．これにより得られた結果をもとに，本提案生成系では，自然な実装よりも効率的な Hadoop MapReduce プログラムを自動生成する．提案生成系を利用すれば，ユーザは，問題を解く逐次プログラムを書くことで Hadoop MapReudce 上で動作する並列プログラムを自動生成できる．

キーワード　大規模データ処理，MapReduce，Hadoop MapReduce，プログラム生成系

Abstract

A Study on Generating Efficient Hadoop MapReduce Programs

Reina Miyazaki

MapReduce is a framework for large-scale data processing proposed by Google, and its open-source implementation, Hadoop MapReduce, is now widely used. Several language systems have been proposed to ease the development of MapReduce programs, for instance, Sawzall, FlumeJava, Pig and Hive. These language systems mainly target applications that can be naturally solved in a MapReduce-like programming model. In this study, we propose a new MapReduce-program generator that accepts programs manipulating one-dimensional arrays.

But, in Hadoop MapReduce, we can develop many kind of programs for the same algorithm due to the many functions and parameters provided. In this paper, first, we focus on the accumulative computation for one-dimensional arrays, and implement and evaluate five programs on Hadoop MapReduce. We discuss what may affect the performance on Hadoop. After that, we generate efficient Hadoop MapReduce programs by using experimental result. By using the proposed generator, users only need to write sequential programs to automatically generate Hadoop MapReduce programs.

key words large-scale data processing, MapReduce, Hadoop MapReduce, program generator