情報処理学会論文誌　プログラミング　Vol. 4　No. 3　94 (June 2011)

発表概要

Hadoop上で動作するSawzallサブセットの実装

中田秀基[†1]　井上辰彦[†1,†2]　工藤智宏[†1]

Sawzallは，Googleが2006年に発表した大容量データの並列バッチ処理に適した言語である．Sawzallの計算モデルはMapReduce型の分散演算であるが，リダクション操作を組み込みのAggregatorに限定することで，エンドユーザによる容易な記述を可能にしている．我々は現在開発中の並列データ処理機構上の言語処理系を開発するための1ステップとして，Scala言語によるSawzall言語のサブセット処理系を実装した．文法やセマンティクスに関しては明確な定義がなかったため，2006年の論文をベースに推測した．その結果，最近公開されたSawzall処理系とは機能的に若干の相違がある．構文解析にScala言語のParser Combinatorを用いることで，処理系の記述量が削減できた．現在の実行対象処理系はHadoopである．Hadoopの Mapper上で言語インタプリタを動作させ，Reducer上では我々の提供するAggregaterを動作させる．Scalaは Java VM上で動作することから，Javaで記述される Hadoop上での実行は容易である．本発表では，本処理系の実装について詳しく述べる．さらに，Hadoopで直接記述した場合と，プログラム量および実行速度の点で比較を行う．比較の結果，プログラム量は大幅に小さくなる一方，実行速度の面でも一定のオーバヘッドがあることが確認された．

An Implementation of Sawzall Subset Interpreter Working on Hadoop

Hidemoto Nakada,[†1] Tatsuhiko Inoue[†1,†2]
and Tomohiro Kudoh[†1]

Sawzall is a script language designed for batch processing of large amount of data, which is introduced by Google in 2006. The processing model of Sawzall is the MapReduce. Sawzall allows programmers only to program 'mappers' to ease the burden. Sawzall provides a set of 'built-in aggregaters', from which programmer choose mapping function. We are developing distributed data processing system for large scaled data. As a part of the project, we have iplemented an interpreter for Sawzall subset in Scala language. We employed paser combinator for lexical parsing. Currently, the system is running on Hadoop. In the paper, we provide detailed implementation of the system. We also evaluated the system with naked Hadoop in terms of program size and execution speed. We confirmed that, with Sawzall, program size is much smaller, while there are certain overhead in terms of execution speed.

（平成23年1月20日発表）

†1 独立行政法人産業技術総合研究所
National Institute of Advanced Industrial Science and Technology
†2 株式会社創夢
SOUM Corporation