

卒業論文試問予稿	Date:	File:	Page:
----------	-------	-------	-------

ビッグデータシステムにおけるバグの実証的研究

現在, ビッグデータシステムがデータマイニング, 情報検索, インターネット上での広告, および購買行動の解析など多岐にわたって利用されている. このようにビッグデータシステムの用途が広がって行くにつれて, システムそのものの信頼性も重要になって来ている. そこで, ビッグデータシステムの信頼性を担保するためには, システムで発生したバグを理解することが重要である. 現在, 様々なソフトウェアシステムに対してバグの理解や修正に関する研究が行われているが, ビッグデータシステムにおけるバグそのものの分類や特性分析に関する研究はほとんど行われていない.

本研究では Apache が公開しているビッグデータを扱う分散処理システム Hadoop と Spark そのもののバグへの取り組みとして, Apache が公開しているバグレポートを集計し, 生成されるバグがどのような特性をもつか調べる. そのことによって, バグが発生した際に分類を同定する足がかりとなる役に立つ. これによって, 今後のビッグデータシステムの発展に対して信頼性を確保していくための導となると考えられる.

研究するにあたって収集したバグレポートは主にタイトル, 重大度, 説明, 作成日, 解決日の情報を元に, 11 の種類に分け, 種類毎のバグの数とその割合や, バグが修正されるまでの期間や, 重大度に応じたバグの個数と割合, バグの種類毎の修正時間とその割合について分析した.

分析の結果, バグの種類に占める上位 3 項目で 46.5%を占めているためにその 3 種類がバグの種類分けをする際の目安になること, 重要度が最大である Blocker には種類毎に現れるか否かに偏りがあることや, すべての種類において半数以上のバグが 3 番目に重大度の高い Major に分類されていることといったビッグデータシステム固有の特徴が判明した.

発 表 者 : 中村 友海
 指導教員 : 趙 建軍 教授
 日 時 : 平成 30 年 2 月 16 日
 11:25~11:40
 場 所 : 302 講義室 (W2-302)