

クラウドソーシングによるデータ列挙処理に関する研究

A Study on the Processing of Crowdsourced Data Enumeration

学籍番号：201221575

氏名：青木 秀人

Hideto AOKI

近年、Web 上の多くのサービスにおいて、クラウドソーシングによるデータ列挙 (Crowdsourced Data Enumeration) が広く行われている。クラウドソーシングによるデータ列挙とは、ネットワークなどを通じて不特定多数の人の力で行う人力検索 (Human-powered Search) の一種であり、特に結果のデータの量が多く、かつ再現率を重視するような検索である。例えば、レストランのレビューサイトにおけるレストラン名の列挙はクラウドソーシングによるデータ列挙である。このように、クラウドソーシングによるデータ列挙は、機械的にデータを列挙することが困難である場合に有用である。

しかし、クラウドソーシングによるデータ列挙には問題がある。それは一般に人力の検索は少数のデータを集める場合には有効であるが、再現率が重要なデータ列挙タスクには適していないことである。なぜならば、列挙されるべきデータが多くなるとワーカは既に入力されているデータの中から未入力 of データや間違いがあるデータを認識する事ができず、結果として、適切なデータの inputs が少なくなり、再現率が低くなるからである。

本研究では、分割統治手法を用いたクラウドソーシングにおけるデータ列挙処理手法を提案する。提案手法では、分割統治手法を用いて列挙範囲を限定した小さなタスクを多数行ってもらうことで、高い再現率の実現を目標とする。提案手法の特徴は、問題分割のプロセスにワーカが参加することにより、列挙範囲を限定したタスクを動的に生成することである。したがって、プログラマがあらかじめこれらのタスクを用意する必要がない。

本論文では、提案手法の説明に加え、提案手法の評価結果について説明する。評価は次の3点で行った。(1) 理論的評価。提案手法が、ある条件の下では単純にデータ列挙を行った場合と同じ結果をもたらすことを示した。(2) 被験者実験。あるデータを対象に単純手法と比較を行い、提案手法のほうが再現率が高くなることを示した。(3) シミュレーション評価。様々な条件でシミュレーションを行い、ワーカのエラー率がよほど高くないかぎり、提案手法が単純手法より一般的に再現率が高くなることを示した。

研究指導教員：森嶋 厚行

副研究指導教員：鈴木 伸崇