「新しい社会システムデザインに向けた情報基盤技術の創出」
2018 年度採択研究者

2018 年度
実績報告書

加藤 誠

京都大学国際高等教育院データ科学イノベーション教育研究センター
特定講師

# オープンデータ利活用のためのデータ検索エンジンの構築

## §1. 研究成果の概要

平成 30 年度はデータ検索エンジンが提供するコンテンツの収集と基本的な検索機能について研究開発を行った.

オープンデータは異なる組織によって公開されているため, 今後新たに公開されるデータを含め網羅的にオープンデータを収集するためには, データが公開されている Web サイトを効率的に発見できる必要がある. そこで, 数ステップ先にデータが存在する可能性があるかを推定し, Web グラフ上に点在するデータを効率的に収集するためのクローラの開発に取り組んだ. このデータ収集用クローラは, 通常の Web クローラと比較した場合, おおよそ 10 倍ほどの効率でデータを発見することが可能である.

クローラの効率化と並行して, Web 上にあるオープンデータの収集を総当たりで実施し, 平成 30 年度時点で 6,069 ドメインから 532,443 件のデータを収集した. この実験によって, 上記のような効率的なクローラの重要性が強調され, また, 今後の実験に必要なデータセットを構築することができた. さらに, 2019 年に収集された 100 億 Web ページ(非圧縮で約 800TB)のデータから, 日本の政府が提供するデータポータルサイト e-Stat にて公開される日本語データの引用を抽出し, 合計 137,388 件を収集することができた. また, 国外のデータポータルサイト(米国の data.gov や英国の data.gov.uk など)中のデータに対する引用 3,604,487 件も収集した. これらのデータは平成 31 年度以降の研究において, データの言語による説明生成などで活用される予定である.

キーワードクエリと属性の対応付けにおいて, 属性名とその属性値を表す言語表現が異なるという問題が存在する. そこで, テキストによって記述される属性を関係分類タスクとして捉え, ゼロショット学習の条件下で属性を同定する問題に取り組んだ.

## §2. 研究実施体制

① 研究者:加藤 誠(京都大学国際高等教育院データ科学イノベーション教育研究センター 特定講師)

② 研究項目

・[データクローラ] オープンデータ Web サイトの発見

・[データクローラ] データ引用の収集

・[データランカ] キーワードクエリと属性の対応付け