「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

2019 年度
実績報告書

加藤 誠

筑波大学図書館情報メディア系
准教授

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

2019 年度は属性オントロジの構築，および，キーワードクエリとデータ属性の対応付け技術の研究開発を行った．

オープンデータ中に含まれる様々な属性に対して，適切な索引を構築するためには，属性間の関係性を理解することが必要になる．例えば，米収穫量に関するデータは「穀物収穫量」というクエリに対して適合する可能性が高いが，厳密な文字列の一致だけを考えると，米収穫量と穀物収穫量は一致していない．これらの属性間に上位下位関係が成立していることがわかっていれば，「穀物収穫量」という語がクエリとして与えられたときに，米収穫量に関するデータも検索結果に含めることが可能になる．同様に，同一関係の判定も検索性能の向上に貢献するはずである．そこで，このような関係にある属性を広く収集するために，Common Crawl にて公開されている大規模 Web コーパスから，同一関係や上位下位関係にある属性の抽出を試みた．より具体的には，2018 年 10 月から 2019 年 4 月の間に収集された約 200 億 Web ページから，数値を含み十分な大きさを持った 5,783,365 個の表を抽出して，これらの表の属性間に成立している関係性を推定し，上記の関係にあるような属性対を抽出した．結果として，19,893 個の数値属性対の同一性と 8,118 個の数値属性対の上位下位関係を発見した．

また，オープンデータを検索するクエリ中には，複合的な属性に基づいた特徴を指す語が含まれる場合がある．例えば，「住みやすい 地域」といったクエリにおいて，「住みやすい」という特徴は地域の「騒音レベル」や「病院数」，「飲食店数」，「公園数」など，地域に関する複数の属性から総合的に判断される．そのため，これらの属性を含むオープンデータが検索意図に対して適合となる可能性が高い．このような検索を実現するためには，検索語と属性間の対応関係を理解する必要がある．そこで，Web 上の順位付きの表データに基づいて，検索語と属性間の対応関係を効果的に学習する方法を提案した．実験では，Web や雑誌から構築された 3 種類のデータセットを用い，提案手法が既存の方法よりも高い精度を達成できることを示した．