研究分担者、協力者募集 !!

論文全文からの引用要素の抽出

2015.07.11 緑川・小野寺ゼミ

天野晃

【背景】現在、挑戦的萌芽研究「実験用マテリアル・リソースの引用にもとづく新しい研究機関影響度指標の開発」において、実験用生物資源の提供情報にもとづく機関インパクト計量に挑戦している。たとえば、"We thank the RIKEN BioResource Center for providing Arabidopsis thaliana T87 cells."のようなセンテンスがあった場合、RIKEN BioResource Centerが提供機関であり、Arabidopsis thaliana T87 cellsが提供される実験用生物資源であることが、知識(辞書)より判明する。この研究では、さらに構文を一般化(thank、providingなどに注目)したパターンで検索し、辞書タームの検索だけでは得られない提供情報(と思われるもの)をも取得する。

【新しい研究で行うこと】挑戦的萌芽研究で行った一般化したパターンでの検索では実験用生物資源以外の提供情報もヒットする期待がある。これらは、すべてを含めて広い意味での引用情報である。新しい研究では、さらにこのパターンを拡張し、あらゆる引用情報の抽出を試みる。このときの引用要素とは、提供者(被引用者)および提供アイテム(被引用アイテム)である。これらは、論文引用で言えば著書とその論文タイトルに相当する。

【方法】網羅的、経験的方法を考えている。第一段階として、センテンスごとに人手により上述のような引用情報を整理する。すなわち、センテンスごとに提供者、提供アイテム、構文内のタームの関係などを整理する。第二段階として、整理した情報より提供者、提供アイテム以外のタームに注目した検索パターンを決定し、計算機により検索を行う。第三段階以降は計算機用検索パターンのチューニングを行う。人手による作業においてすでに正解集合は得られているので、この結果に近づけるような検索パターンを模索していく。最終的には提供者-提供アイテムのデータセットとして公開する。

【対象】PMCを予定している。

【応募時期と応募対象】今年度募集の科研基盤Aに応募予定。