

メタデータのグラフ構造を利用した
メタデータタームの類似度算出手法
A Similarity Calculation Method Between Metadata Terms
Using Graph Structure of Metadata

学籍番号：201721665

氏名：金城 良大

Kinjo Ryota

新しいデータ公開の仕組みとして Linked Open Data(LOD) が注目を集めている。LOD とは標準化されたデータ表現とアクセス方式を持つ、相互運用性の向上を目指したデータである。2013 年の G8 サミットでは LOD として政府データを公開することを推奨する”オープンデータ憲章”が合意されるなど、LOD はデータ公開の標準として期待をされている。しかし LOD は研究者の間で普及が進む一方、専門知識を持たない人には扱いが難しい。本研究では LOD の扱いの難しさのうち、データ記述に使用するメタデータタームの探索と用法理解の難しさに注目した。メタデータターム(ターム)とは記述する事物のデータ項目であるプロパティと記述する事物の属性を表すクラスの総称である。タームは利用方法や意味解釈のための情報が形式的に定義され、その定義に従って利用されることが望まれている。タームの探索と用法理解を難しくしている原因は、語彙定義が適切になされていないことや参照できないことが考えられる。そこでタームの類似関係がわかると探索と用法理解に役立てることが期待できる。例えば探索において、探索者がメタデータ記述に利用するための候補としているタームと類似関係にあるターム群を参照することができれば、より目的に即したタームの発見を支援できる。

本研究ではメタデータタームの探索と用法理解を支援するためにターム間の類似度算出手法を提案した。語彙定義がない状況でも適用できるようにインスタンスの局所的なグラフ構造の一致からからその類似度の算出を行った。具体的には 2 つの異なる LOD を結合し、グラフを構成するノードの表現ベクトルを求めるグラフエンベディング手法を適用する。そして、それぞれのノードの表現ベクトルを取得して、そのコサイン類似度の算出を行った。また評価用データセットでの実験を通じて、LOD をそのまま入力とするのではなく事前にノード間のマッチング、プロパティのエッジ化、ノード間の関係に応じたエッジの重み付けなどの工夫を行うことで算出される類似度が改善されることを確認した。

研究指導教員：杉本 重雄

副研究指導教員：永森 光晴