Linked Open Data 環境における
メタデータ記述語彙の類似度算出手法
A Calculation Method of Similarity between Metadata Terms in Linked Open Data Environment

学籍番号：201521635
氏名：二十歩　亮介
Ryosuke NIJUBU

現在，様々なデータ形式のデータが Web 上に公開されている．なかでも相互運用性と発見性に優れた Linked Open Data（LOD）環境での公開が望まれている．そのための標準としてしばしば Resource Description Framework（RDF）が利用される．RDF ではデータの記述にメタデータ語彙を使用する．メタデータ語彙にはデータ項目を記述するためのプロパティと記述対象自身の分類を記述するためのクラスが含まれ，プロパティとクラスを総称してタームという．データ記述の際には既存のメタデータ語彙から適切なタームを選択し再利用することで作成されたデータの相互運用性を高めることができる．タームの探索は Linked Open Vocabularies（LOV）に代表されるメタデータ語彙探索支援システムを用いて行われることが多い．それらのシステムが提供する検索環境を利用することで利用目的に応じたタームの候補を発見できるが，適切なタームが候補に含まれているとは限らない．そこで，利用目的に完全に合致せずとも類似するタームから関連するタームを参照することができれば，適切なタームの発見を支援することができるのではないかと考えた．本研究では関連するタームを参照可能にするためのタームの類似度算出手法を提案する．

本手法はタームの名前と説明文からタームを特徴付ける単語である代表語を抽出し，それらの分散表現を比較することでタームの意味に基づく類似度を算出する．本研究では LOV に登録されているメタデータ語彙のタームを対象に類似度を算出可能にした．独自に作成した X-ABC 評価セットを用いて，類似度算出における設定を変えて評価実験を行い，結果を比較したところ，多数の代表語を利用するよりもより重要な少数の代表語を利用することで類似度算出の精度が向上することがわかった．また，代表語として抽出してもベクトル化できない単語が多く存在することがわかり，代表語の抽出と選出方法ならびに Word2Vec に与える学習データの改良の必要性に関する知見を得た．

研究指導教員：杉本　重雄
副研究指導教員：永森　光晴