Linked Open Data におけるデータセット間のリンク支援手法

新井　叡樹

Linked Open Data（以下 LOD）は他のデータセットへリンクすることによって利便性を向上させることができるため, LOD においてデータセット間のリンクは重要である. しかし, 異なるデータセット同士をリンクするには，互いのデータセットに含まれる個々のデータを逐一確認する必要があり，その作業量は少なくない．そこで本研究では，LOD におけるデータセット間のリンク作業を支援する手法を開発することを目的とする.

LOD のデータセット間のリンクを自動で付与する手法がいくつか提案されているが，現状, 機械的な処理だけでは意味を考慮した正確なリンクを付与することは困難である. そこで, 人手による効率的なリンク付与を実現するために, 本研究では機械的に抽出したリンク候補の判定作業について，マイクロタスク型クラウドソーシングを適用する手法を提案する. 本研究では, それぞれのデータセットで記述されているリンク候補についての情報を表形式で表し，表の内容を比較しリンク判定をすることをタスクとする．また，LOD の多くは RDF で記述されており，RDF ではデータを主語・述語・目的語の三つ組で表す．何らかの事物をリソース，主語と目的語の関係をプロパティ，文字列や数値をリテラルと呼び，主語はリソース, 述語はプロパティで記述され, 目的語はリソースで記述される場合とリテラルで記述される場合がある．また，リソースとプロパティは URI で識別される．本研究では，データセットにおいて記述されているリソースの内主語であるものをリンクの対象とし，リンク方法は同じ事物を示すリソース同士をリンクするときに用いられる owl:sameAs を用いる. 候補抽出では，各データセット間のリソース同士で類似度を求め，それが閾値を超えたらリンク候補とする. リソース同士の類似度は, 比較するリソースを主語とするプロパティの目的語の内リテラルであるものの重みを算出し，その重みと比較するリソース間のリテラルの類似度の 2 つの値から算出する. 次に，リンク候補に関する情報を表形式に変換する処理では，リソースの名前を表すリテラルを表の一行目，二行目にリソースの URI, 三行目からは一列目にプロパティ，二列目にプロパティの目的語を配置した表を作成する. また, 視認性とタスクの作業効率を向上させるために，リソースとプロパティについては, URI をそのまま表示するのではなく，URI を短縮形に変換したものを表示する.

本研究では，提案手法で作成した表がリンク判定のタスクとして適切であるかを検証するために筑波大学の学生 12 人に実際のデータセットから作成したタスクを行ってもらい, その見やすさについてのアンケートに答えてもらった. その結果, タスクの正答率は 95%, 作成した表がタスクにおいて見やすいと回答した人は 7 人，LOD 一般に関する知識とタスクの正答率の相関係数は 0.085 となった．以上より，提案手法で生成したタスクはデータセット間のリンクを判定する際に有効であると考えられる.

（指導教員　阪口哲男）