LOD を参照しているドキュメントを利用した検索支援

藤原　暉

LOD（Linked Open Data）は、Web 上でデータを公開・共有する際の仕組みとして、URI による識別子の付与と RDF に基づき構造化されたデータの記述を特徴とし、データの相互運用性を高める重要な技術基盤である。LOD 利用における現時点の課題の一つとして、LOD データセットの網羅的かつ最新情報に基づいた検索環境の整備が挙げられる。従来から用いられているデータカタログサイトの手動更新に依存する仕組みでは情報鮮度と網羅性が低下するという課題があり、研究では、その課題を克服した LOD データセット検索システム構築のための、LOD データセットと Web ページ間の参照関係を活用した関連情報収集とキーワード抽出手法を提案する。

本研究で提案する手法は、まず先行研究に基づいた SPARQL Endpoint の自動収集モジュールにより SPARQL Endpoint の URL を収集する。次に、収集した URL をリンクしている Web ページを検索し、それらをデータセットの「関連ドキュメント」として収集する。本研究で収集した 58 件の SPARQL Endpoint を対象に分析し、Web ページ上でデータセットの説明文やメタデータが記述される HTML 構造パターンを特定し、それに基づくスクレイピングを行った。最後に、収集した関連ドキュメントから、収集元の URL と SPARQL Endpoint の URL の類似度、文書構造や出現頻度、周辺文脈などの特徴を考慮した解析と tf-idf による重み付けを経て、重要度の付与されたキーワード抽出を行う。

評価実験では、サンプルとしての SPARQL Endpoint を集めた上での関連ドキュメントの収集において、少なくとも５件以上の Web ページから関連ドキュメントを得ることができた SPARQL Endpoint の割合が 89.5%であり、この手法によるキーワード抽出の一定的な安定性を確認できた。加えて、単純な抽出に対しての HTML のパターン分けによる抽出量の優位性も確認できた。一方で、HTML パターンの更なる細分化や、動的生成コンテンツを含むより多様な種類の Web ページ上の情報や文脈を考慮しての、関連ドキュメントの収集方法の改善が今後の課題として残された。

本研究の成果は、より多くのデータセットを網羅した効果的な LOD データセット検索支援システムの実現に向けて大いに役立つことが期待される。そして、提案手法をより洗練することによって得られるキーワード群は、データセットの特徴をより正確に表現し、ユーザの検索意図に即した検索結果の提供に導くことが期待できる。

（指導教員　阪口 哲男）