LODにおけるリンク候補の推薦

関口 賢

Linked DataはURIを用いて他のデータとリンクされているデータであり，誰でも自由に利用できるようにオープンライセンスで公開されたデータがOpen Dataである．この2つを組み合わせたものがLinked Open Data(以下LOD)である．LODはResource Description Framework(以下RDF)によって記述されることが多い．近年，行政機関等の組織が所有する様々なデータをLODとして公開し，利活用することが盛んになってきている．LOD化の際には，リンク元とリンク先のデータセットのデータを逐一確認し，リンク元データの一つ一つに適切なリンク付けを行う必要がある．その際，リンク先候補を人手で探すのはコストが高い．そこで本研究では，LODにおけるデータセット間のリンク付けを支援するために，リンク先候補の推薦を効率よく行うことを目的とする．

RDFは，主語・述語・目的語によって構成されるトリプルの組み合わせでデータを表現し，主語と述語はURI，目的語はURI又はリテラルによって記述される．主語から名前などの文字列が目的語としてリンクされていることが多い．そこで本研究では，文字列の類似度を編集距離に基づいて計算し，その類似度の高いものから順にリンク先データ候補とする．また，編集距離をリンク元とリンク先のデータセットの各データについて総当たりで求めるのはコストが高いため，リンク先データについてシグネチャをキーとした索引付けを行う．リンク元データのシグネチャと索引の各キーの排他的論理和演算を行い，得られた値で1となるビットが最も少ないシグネチャを持つリンク先データと編集距離を計算する．得られた編集距離の短い順に指定件数番目までリンク先データ候補とする．

本研究では，リンク元に2つ，リンク先に6つのデータセットを選び，その中の住所記述を用いて評価実験を行った．リンク先データ候補を推薦する際にシグネチャのビット長を変化させてCPU時間を計測し，その際の推薦の精度を評価した．推薦された候補上位3件に正解が含まれているかどうかで精度の評価を行った．なお，今回用いたデータではリンク元と一致するものが含まれているので，それが候補に含まれている場合に正解とした．索引を用いない総当たりの場合，精度は100%となり，編集距離に基づく類似度を用いる手法が期待通りであることがわかった．索引を用いた場合，総当たりよりも最大で約125倍の処理速度向上が見られた．しかし，精度は最高でシグネチャのビット長を64とした際の約84.6%，最低でビット長を8とした際の約10%であった．この結果より，索引付けを行うことで処理速度を向上させることに成功したが，推薦精度に課題が残された．この改善には対象データに合わせたシグネチャの計算法の検討等が必要と思われる．

（指導教員　阪口哲男）