

Linked Open Data に対するキーワード検索手法の提案

奥村 彩水[†] 天笠 俊之[‡] 北川 博之[‡]

[†] 筑波大学情報学群情報科学類 [‡] 筑波大学システム情報系

1 はじめに

Linked Open Data (LOD) とは、相互にリンクされたオープンデータのネットワーク (Web of Open Data) を構築し、機械処理可能なデータを広く公開しようとする試みである。近年のデータ公開の流れと相まって、政府や企業等がデータを公開する手段として注目されている。LOD において、データは Resource Description Framework (RDF) によって記述され、URI により相互に参照される。また利用者は、クエリ言語 SPARQL^{*1} を利用することで、問合せを行うことができる。

しかしながら、SPARQL 問合せを記述するために、利用者は SPARQL 言語そのものを習得することが必要である。さらに、問合せ対象となる LOD データそのものの構造を知る必要がある。LOD データは一般に複雑なグラフ構造を持つため、複雑かつ巨大なデータの場合、特に後者は困難である。また、SPARQL の結果はランキングされていないため、重要度の高い結果を判別することが困難である。

この問題に対して、本稿では LOD に対するキーワード検索手法を提案する。関連研究において行われているような、RDF トリプルを単位とする検索ではなく、エンティティをベースとし、ObjectRank[1] によるランキングを行う点に特徴がある。

2 関連研究

Lei ら [2] は、セマンティックウェブに対するサーチエンジンとして、Google ライクなインタフェースを使った検索が可能なキーワード検索エンジンを実現するシステムを構築している。一方、一瀬ら [3] は、LOD の中核データである DBpedia^{*2} に対する SPARQL 検索結果のランキング方法として、PageRank アルゴリズムを用いた手法を提案している。一般の LOD データに対して、エンティティをベースとして、ObjectRank によるランキングを行う手法はこれまで提案されていない。

A keyword search technique for Linked Open Data

Sayami Okumura[†], Toshiyuki Amagasa[‡] and Hiroyuki Kitagawa[‡]

[†] College of Information Science, University of Tsukuba

[‡] Faculty of Engineering, Information and Systems, University of Tsukuba

^{*1} <http://www.w3.org/TR/rdf-sparql-query/>

^{*2} <http://dbpedia.org/>

3 ObjectRank

ObjectRank[1] は、PageRank を拡張したグラフに対するノードのランキング手法である。PageRank とは異なり、複数種類のノードやエッジを扱うことができる。まず、ノードおよびエッジの種類とエッジの重みを定義した Authority Transfer Schema Graph (SG) を構成する。次に、SG に基づき、分析対象のグラフから Authority Transfer Data Graph (DG) を構築する。DG に基づいた行列を A とすると、ObjectRank による評価値 $\mathbf{r} = [r(v_1), \dots, r(v_n)]^T$ は以下の式で求められる。

$$\mathbf{r} = d\mathbf{A}\mathbf{r} + \frac{1-d}{|V|}\mathbf{e}$$

ここで d はダンピングファクタ、 $\mathbf{e}$ は全ての要素が 1 の n 次元列ベクトルである。実際は、上で求めた global ObjectRank に加えて、検索キーワードにマッチしたノードに関する personalized ObjectRank を求め、両者を重み付き統合することで ObjectRank 値を求める。

4 提案手法

4.1 エンティティに基づく LOD のキーワード検索

LOD (RDF) データはトリプル $t = (s, p, o)$ の集合 T からなる。ここで s, p, o はそれぞれ主語、述語、目的語である。 T に対するキーワード検索とは、ユーザによって与えられた検索キーワード集合 $K = \{k_1, k_2, \dots, k_m\}$ に対して、適合するエンティティをランク付きで返す処理である。

まず、LOD データにおける検索対象となるエンティティを与える。一般に LOD データには、多様なエンティティが含まれる。本手法では、システム構築者等によって事前に検索対象となるエンティティが指定されているものとする。それには、クラスによる指定、SPARQL 問合せによる指定等を利用する。

検索対象となるエンティティが決まると、次に各エンティティに対応するサブグラフの抽出および特徴抽出を行う。まず、各検索対象エンティティの URI u_i を主語として持ち、リテラルを目的語に持つトリプル $(u_i, *, l_j)$ を探索する。これをエンティティサブグラフ (ESG) と呼ぶこととする。次に、1) 各エンティティ (に対応する URI) u_i 中に出現する単語、および、2) リテラル l_j から単語抽出を行い、これを u_i の特徴語とする。本研究では、この ESG を対象としてキーワード検索を行う。

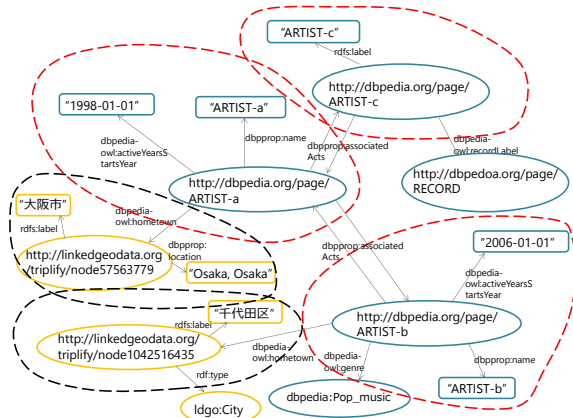


図1 エンティティサブグラフ (ESG).

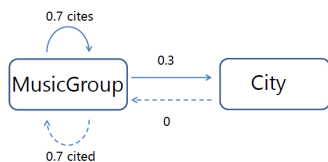


図 2 スキーマグラフ (SG).

例を使って説明する(図1)。この例では、MusicGroup および City が検索対象クラスとして指定され、その結果それぞれ三つ(赤点線内)と二つ(黒点線内)の ESG が抽出されている。

4.2 スキーマグラフ (SG) の作成

ObjectRank によるランキングを実現するためには、ノードおよび（重み付き）エッジの種類を規定した Authority Transfer Schema Graph (SG) を作成する必要がある。このため、検索対象クラスをそれぞれ SG におけるノードとする。さらに、LOD データ中に、そのクラス間にエッジがある（トリプルが存在する）場合、SG 中のノード間に対応するエッジを作成する。この際、エッジの重みは、LOD データ中に出現するトリプルの統計量等を使って導出する。また、システム構築者等による重みの補正、あるいは直接的な指定等も考慮する。

例として、図 1 の場合、**MusicGroup** および **City** に対応するノードが作成され、対応するエッジが作成される。なお、**MusicGroup** 同士の参照は自己ループとして表現され、**ObjectRank** の定義に従い、片方向のエッジには必ず逆向きのエッジが作成される。また、エッジの重みは、データの性質を考慮して適当に設定しているが、適切な重みの導出は今後の課題である。

4.3 データグラフ (DG) の作成

ObjectRank の解析の対象となるグラフ (Authority Transfer Data Graph; DG) は, SG および ESG を元に作成する. 具体的には, ESG の各エンティティをノード集合として, LOD データから誘導される誘導部分グラフを用いる. このとき, 各エッジの重みは, ObjectRank のアルゴリズムに従って SG 中のエッジの重みを元に算

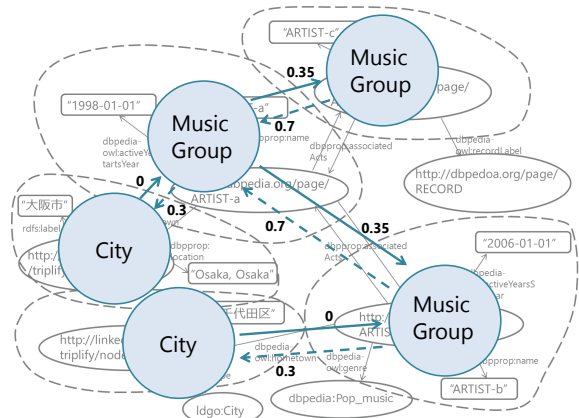


図 3 データグラフ (DG).

出される.

図 3 に図 1 から導出される DG の例を示す。DG におけるエッジの重みは、そのエッジに付与されている重みを、エッジの元ノードの持つ出次数で割った値となる。例えば、MusicGroup の自己参照エッジの重みは 0.7 である。このため、DG においては、<http://dbpedia.org/resource/ARTIST-a> から出力されるエッジの重みは、その出次数である 2 で割った 0.35 となる。

得られた DG を元に，3 節で説明した ObjectRank の計算式を適用することで，各エンティティに対する評価値を計算することができる．これにより，検索結果のランキングを行う．

5 まとめ

本稿では、LOD においてエンティティを対象としたキーワード検索、および、ObjectRank を利用した LOD に対する検索結果のランキングを行う手法を提案した。今後は、実データによる評価を行うとともに、スキーマグラフにおけるエッジの重みの自動調整について検討する予定である。

謝辭

本研究の一部は、共同研究費（富士通研究所），文科省“実社会ビックデータ利活用のためのデータ統合・解析技術の研究開発”，および，科研費（25240014）による。

参考文献

- [1] Balmin, A. Hristidis, V. Papakonstantinou, and Y. ObjectRank:Authority-based keyword search in databases. VLDB 2004.
- [2] Yuanguai Lei, Victoria Uren, and Enrico Motta. Semsearch:A Search Engine for the Semantic Web. Managing Knowledge in a World of Networks.
- [3] Shiori Ichinose, Ichiro Kobayashi, Michiaki Iwazume, and Kouji Tanaka. A Ranking Method for DBpedia Resources based on Retrieval Results with SPARQL Queries. JSAI 2013.