1K-05

Linked Open Data における確率モデルに基づいたリソース間の経路ランキング手法

西田寛章† 古崎晃司† 駒谷和範†
大阪大学産業科学研究所†

1 はじめに

Linked Data とは構造化されたデータをリンク付けして, それらを公開できる仕組みを提供する実践的方法であり, それに従った Open Data を Linked Open Data(LOD)と呼ぶ. 一般的に Linked Data で用いられるデータモデルにおいて, Linked Data はラベル付有向グラフで表される. また, そのノードをリソース, ラベルをプロパティ, 始終点リソースとプロパティの 3 つの組をトリプルと呼ぶ. ここで, 有向グラフにおいての頂点がリソース, 辺がトリプルに対応する.

Linked Data の原則[1]によれば, リソースの URI にアクセスされた際には有用な情報を標準的なフォーマットで提供するべきとされている. しかし, 「有用な情報」については明確に決まっておらず, 多くのシステムはそのリソースに直接繋がっている全てのトリプルを列挙している. これは順序付けされていないために見辛く, 加えて直接繋がっていない重要なリソースが列挙されないという問題がある.

2 確率遷移モデルに基づく経路ランキング

本研究では, 「有用な情報」として, 指定したリソースから別リソースまでの関連度を表すスコアを計算して 降順にランキングし, その経路とともに提示する手法を提案する.

最初に, LOD と対応する有向グラフの頂点を状態, 辺を遷移とみなしたマルコフ連鎖上で生起確率の大きな経路を列挙し, その後列挙した経路を元にスコアを計算する.

2.1 遷移確率の割当

あるリソースAから直接繋がっているリソースBへの遷移確率は, Bの重要度に比例すると仮定すると, AからBへの遷移確率は式(2.1)となる.

$$P(A,B) = \frac{Importance(B)}{\sum_{v \in Ao} Importance(v)} \quad \cdots (2.1)$$

ここで, $P(A,B)$はAからBへの遷移確率, $Importance(v)$はリソースvの重要度, AoはAと直接繋がっているリソースの集合である. ここで, 明らかに$B \in Ao$となる.

リソースの重要度の指標について, 一瀬ら[2]は PageRank を用いる手法を提案している. これを参考にvの PageRank を$Importance(v)$とした.

2.2 経路問題への帰着

節 2.1 に定義した確率遷移上で, あるリソースAを始点として, 任意のリソースへの確率の大きな k 本の経路の経路を列挙する. これは非負辺制約付きの k 最短経路問題と等価であり, Eppstein の手法[3]を用いれば$O(E + V\log V + VK)$で求める事ができる. ここで, Vはリソース数, Eは辺の数である.

Aから任意のリソースまでの経路を得た後, 式(2.2)に従って経路のスコアとリソースのスコアを定義する.

$$score(A,v) = \sum_{path \in AtoV} score(path) \quad \cdots (2.2)$$

ここで, $score(path)$は経路のスコア, $AtoV$はAからvに至る上位 k 本の経路の集合, $score(A,v)$はAからvまでのスコアを表す. また, $score(path)$は経路の生起確率$P(path)$とする.

3 経路のスコア改善

2 章で述べた単純な遷移確率に基づく手法をベースライン手法として, その改善案を提案する. 表 1 にそれぞれの手法で得られたランキングを示す（詳細は 4 章で述べる）.

3.1 ハブ検出による改善

リソース「大阪府」を始点として 2 章で述べたベースライン手法を用いた結果には, 関連が高くないと思われる「日本郵便」などのリソースへの経路が高くスコア付けされる問題が見られた（表 1 ベースライン手法）. ここで, 「大阪府」から「日本郵便」までの経路は, 多くの場合「日本」を経由している. そのため「日本郵便」は「日本」に強く依存してスコア付けされていると考えられる. このような多くの経路に含まれているリソースをハブと呼ぶことにする.

ここで, ハブを含んだ経路は関連が低いと仮定し, ランキングから削除する. Aからvに至る経路がハブを含むかを判定する式を式(3.1)に示す.

$$\max_u \left\{ \frac{\sum_{path \in paths(AtoV,u)} score(path)}{score(A,v)} \mid u \in S \right\} < X \quad \cdots (3.1)$$

ここで, Sは$AtoV$の中に含まれるリソースの集合, $paths(AtoV,u)$は$AtoV$のうちリソースuを含む経路の集合, Xは閾値を表し, Xは適当な定数とする.

3.2 類似経路検出による改善

節 3.1 で述べたハブ検出手法を用いて「大阪府」からリソースをランキングした時, 「ABO 式血液型」などの重要度の低いと思われるリソースが高くスコ

A ranking method for paths among resources in Linked Open Data based on stochastic model
†Kansho NISHIDA, Kouji KOZAKI, Kazunori KOMATANI
†Department of Knowledge Science, I.S.I.R, Osaka University

ア付けされる問題が見られた（表1ハブ検出ありの列を参照）．これは「大阪府」にリンクしている人物の多くが「ABO 式血液型」にもリンクしているためである．

ここで, ある経路のスコアについて再考する．「大阪府」から「ABO 式血液型」までの経路の多くは次の様になっている．

・「大阪府」-(location)- <Person> -(bloodType)- 「ABO 式血液型」

（ここで，“「 」”はリソース，“-( )-”はリソース間に存在するトリプルのプロパティ，“< >”はリソースのクラスを表す．）

ここで, 似ているパスが多くある場合，1つあたりの重要度は低下すると仮定する．ここでは簡単にプロパティのみに着目し, (location, bloodType)とのように経路に含まれる全てのプロパティが順序を含めて一致する時, 経路が類似するとする．そこで, 経路のスコアを式(3.2)に従って再定義する．

$$score(path) = \frac{P(path)}{count(path)} \cdots (3.2)$$

ここで,$path$はvに至る k 本の経路の 1 つであり, $count(path)$は k 本の経路のうち$path$と類似する経路の数である．

4 実験条件

2 章で述べた手法，3 章で述べた 2 つの改善手法によるランキング結果を比較した．実験データにはDBpedia 日本語版のデータセット[4]を用いた．リソース数は 525709, トリプル数は 2991725 であった．

PageRank の計算は一瀬らに従い, Eppstein の手法で用いる経路の数は 10 とし, ハブ検出で用いる閾値 X は 0.8 とした．

以上の条件で「大阪府」からそれぞれの手法でランキングした．最も計算時間が掛かる箇所は Eppstein の手法の$O(E + V\log V + VK)$であるが, V = 525709, E = 2991725, K = 10 であり, クロック数 2GHz の CPU, 8GB の RAM を搭載した計算機で数秒で終了する程度であった．

表 1　「大阪府」から到達する上位リソース

ベースライン手法		ハブ検出あり		ハブ/類似経路検出あり	
4	日本郵便	8	株式会社_(日本)	8	兵庫県
5	東京都	9	株式会社	14	京都府
10	北海道	10	兵庫県	28	三重県
16	埼玉県	11	J-POP	41	滋賀県
17	講談社	16	ABO 式血液型	44	和歌山県
18	東日本旅客鉄道	18	京都府	45	J-POP
19	集英社	20	代表取締役	47	奈良県
22	EMI ミュージック・ジャパン	28	ロック_(音楽)	48	株式会社
23	兵庫県	33	三重県	51	ABO 式血液型
25	ポニーキャニオン	40	作曲家	58	アップフロントワークス

表 2　「大阪府」から「兵庫県」へ至る上位経路

大阪府 -(country)- 日本 -(country)- 兵庫県
大阪府 -(location)- 大阪市 -(neighboringMunicipality)- 兵庫県
大阪府 -(routeStart)- 国道 2 号 -(city)- 兵庫県
大阪府 -(city)- 国道 176 号 -(city)- 兵庫県
大阪府 -(neighboringMunicipality)- 和歌山市 -(neighboringMunicipality)- 兵庫県
大阪府 -(routeEnd)- 国道 477 号 -(city)- 兵庫県
大阪府 -(campus)- 関西学院大学 -(locationCity)- 兵庫県
大阪府 -(neighboringMunicipality)- 尼崎市-(location)- 兵庫県
大阪府 -(locationCity)- 西日本旅客鉄道 -(locationCountry)- 日本 -(country)- 兵庫県
大阪府 -(city)- 国道 171 号 -(city)- 兵庫県

5 実験結果と考察

本研究で提案した手法により, 従来手法では得られなかった直接繋がっていないリソースまでの経路も含めランキングすることが出来た．

表 1 に, 最も短い経路長が 2 以上である上位 10 件のリソースと, 2 未満であるリソースを含めた時の順位を示す．

ハブ検出ありの手法は, ベースライン手法に比べて「日本郵便」など, 「大阪府」と関連が低いと思われるリソースがランキングから除外されていることが確認できる．

さらに類似経路検出も加えた手法は, ハブ検出のみの手法に比べて「株式会社」などの関連が低いと思われるリソースの順位が下がっている一方, 「兵庫県」などの順位が上昇しており, 経路の類似性を加味する事が有効であることを示している．

表 2 に, 取得される経路の具体例として, ハブ検出/類似経路検出ありの「大阪府」から「兵庫県」までの k（=10）件の全ての経路を示す．

リソースの重要度に関しては, 奥村ら[5]は PageRank の発展手法である ObjectRank を用いることを提案している．また DBpedia の場合, 元データとなっている Wikipedia の閲覧数の集計が公開されており[6], これを用いるのも一考の余地がある．

類似経路検出に関して, 今回はプロパティだけを用いた．これに加えてリソースの type や property の subProperty などを用いることで改善の余地があると思われる．

参考文献

[1]Tim Berners-Lee. Linked Data - Design issues, 2016. https://www.w3.org/DesignIssues /LinkedData.html. 7, 26, 82

[2]一瀬詩織, 小林一郎, 岩爪道昭, 田中康司: DBpedia を対象にしたリソースのランキング手法における一考察, 情報処理学会 第 75 回全国大会講演論文 Vol.2013, No.1, pp.601-603(2013).

[3]Eppstein, D. Finding the k Shortest Paths. SIAM Journal on Computing, Vol.28, No.2, pp.652-673.

[4]http://wiki.dbpedia.org/downloads-2016-04

[5]奥村彩水, 天笠俊之, 北川博之:Linked Open Data に対するキーワード検索手法の提案, 情報処理学会第 77 回全国大会講演論文集 Vol.2015, No.1, pp.683-684(2015).

[6]https://dumps.wikimedia.org/other/analytics/