5N-4

プロパティに着目したSPARQL問い合わせ結果のランキング手法

一瀬詩織 † 小林一郎 † 岩爪道昭 ‡ 田中康司 ‡
† お茶の水女子大学大学院人間文化創成科学研究科理学専攻情報科学コース
‡ 独立行政法人情報通信機構

1 はじめに

近年データの共有・再生産を目的として，RDF 形式を採用したデータセットが Web 上に活発に公開されている．これらの構造化されたデータセットからデータを抽出する手法としてクエリ言語である SPARQL を用いる手法があるが，SPARQL を用いた問い合わせでは検索結果の量が多い場合，どの結果が検索者にとって有用であるかが分かりづらい．本研究では RDF データセットの問い合わせに用いられる SPARQL 言語を対象とし，クエリ構造を利用した検索結果のランキング手法を提案する．結果を重要度の順にランキングすることで SPARQL クエリ検索の支援を行う．

2 関連研究

本研究での提案手法と同じく，Semantic Web データの構造化情報を利用しクエリ検索結果のランキングを行う手法には，Bamba ら [1]，Mulay ら [2] の研究がある．Bamba らの研究 [1] では HITS らのアルゴリズムを参考にしたアルゴリズムを用いてリソースの重み付けを行い，さらにクエリ検索結果のグラフ構造を用いて結果のスコアを計算する手法を提案している．Mulay らの手法 [2] ではデータセット，リソース，トリプルの三層構造を用いたリンク構造の解析により，リソースの重み付けを行っている．ただし利用している情報が SameAs リンクと関連した情報に限定されており，それ以外の様々なプロパティ情報は考慮されていない．

本研究ではデータベース内のリンク情報からリソースやプロパティの重要度を定義し，Bamba らの研究におけるクエリ検索結果の評価アルゴリズムに RDF トリプルの評価式を導入したアルゴリズムにより，SPARQL クエリ検索結のランキングを行う手法を提案する．

3 提案手法

SPARQL クエリによって得られた検索結果をそれぞれクエリ構造に基づくグラフとみなし，それぞれのグラフについて本手法のアルゴリズムを用いた評価を行う．グラフには複数のリソースとプロパティが含まれており，アルゴリズムによってそれぞれの重要度を重み付けすることでグラフ全体を評価する．全てのグラフを評価し，評価値に基づいて決定された順位を検索結果のランキングとして出力する．

3.1 リソースとプロパティの重要度

リソースの重要度には情報検索の分野で利用される PageRank アルゴリズム [3] を用いる．またプロパティの重要度計算にはプロパティの主語が属するクラスに着目し，あるクラスにおけるプロパティの出現頻度とクラスに対する希少性を考慮した指標 PF・ICF を定義する．プロパティ頻度 (PF) はあるクラスにおいて，そのプロパティが使われる頻度を表す．逆クラス頻度 (ICF) はすべてのクラスにおいてそのプロパティが使われる頻度の逆数を表す．この指標は文書処理の分野で用いられる TF・IDF を参考にしたもので，作家や大学などの特定のクラスにおいて，多く出現するプロパティに高い値を与える．

3.2 PFICF を用いたリソース評価手法（提案手法１）

リソースをノード，プロパティをエッジとし，検索結果から作成したグラフに Bamba ら [1] のクエリ評価アルゴリズムを適用してグラフの評価を行う．ノードとエッジの重要度には 3.1 で定義した指標を用いる．

3.3 RDF トリプルを単位とした評価手法（提案手法２）

RDF データの情報はプロパティとリソースの３つ組 (トリプル) により記述される．Bamba ら [1] における結果グラフの評価アルゴリズムではグラフ中のリソース (ノード) とプロパティ(エッジ) を別々に評価しているが，本研究ではより RDF データに適した評価を行

A Ranking Method for the Results of SPARQL Query focusing on RDF Properties

†Shiori ICHINOSE(ichinose.shiori@is.ocha.ac.jp),
†Ichiro KOBAYASHI(koba@is.ocha.ac.jp)
‡Michiaki IWAZUME(iwazume@nict.go.jp)
‡Kouji TANAKA(tanaka@nict.go.jp)
†Advanced Sciences, Graduate School of Humanities and Sciences, Ochanomizu University,2-1-1 Ohtsuka Bunkyo-ku Tokyo 112-8610
‡National Institute of Information and Communications Technology, 3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0289

うため，ノードとエッジを同時に評価するトリプルの評価式をアルゴリズムに導入する．

RDF トリプルの主語，述語，目的語を n_s，e，n_o で表し，評価式 $TripleScore$ を以下のように定義する．

$$TripleScore = \frac{Imp(n_s) \times PFICF(n_s, e) \times Imp(n_o)}{linkNum(n_s) + linkNum(n_o) - 1} \quad (1)$$

$Imp(n)$，$PFICF(n, e)$ はそれぞれリソースの重要度，プロパティの重要度を表す．また，$linkNum(n)$ はノード n から出るエッジの本数を表す．ここで $linkNum(n)$ によるスコアの分割は，グラフ評価アルゴリズムにおいて同じリソースを主語として，目的語としてなど複数回評価する場合に行われる．

手法２におけるグラフ評価のアルゴリズムを以下に示す．ここで $decayFactor$ はユーザの興味の強さの減退を表す定数である．1.0 以下の値に設定することで SELECT 節で選ばれたリソースやプロパティとグラフ上の距離が遠いトリプルの重要度を低減させる．

アルゴリズム:

1. $decay = 1.0$, $score = 0.0$ に初期化する．
2. Adj を SELECT 節で選択されたノードを含んだトリプルの集合で初期化する．
3. Adj が空になるまで以下を繰り返す:
   (a) $ClassedEdges$ を Adj のノードのクラスとノードから伸びたエッジの集合とし，(c, e) で表す．
   (b) $score(r) += \sum_{t \in Adj} TripleScore[t] * decay$
   (c) $decay *= decayFactor$ $(decayFactor < 1.0)$
   (d) Adj と隣接した，まだ訪れていないトリプルで Adj を初期化する．

4 実験

4.1 実験仕様

SPARQL の使用目的として,「条件を指定してリソースを取得する場合」と「リソースに関するトリプル (情報) を取得する場合」の２つのケースを想定した．それぞれ「クラス University に属するリソース」「クラス City に属するリソース」,「京都に関する情報」「東京大学に関する情報」の 4 つのクエリを用いて結果を取得し，ベースライン，提案手法のそれぞれを用いたランキングを行った．ベースラインには Bamba ら [1] の論文にて提案されている手法を用いたが，再現が困難なリソースの重要度評価の部分には本手法と同様の PageRank アルゴリズムを利用している．$decayFactor$ は両手法とも 0.5 に設定した．

各提案手法とベースラインとの有意差を検証するため，各クエリにおけるランキング結果上位 20 件について 12 名の被験者に１～５の評価を行ってもらい，４以上の評価を「良い」，それより低い評価を「悪い」とした評価の２クラス分類を行った．有意差検定には Fisher の直接確率検定を用い，有意水準 $p < 0.01$ をもって有意差があるとして検定を行った．

4.2 結果と考察

スコアの平均値を表 1 に示す．また Fisher 検定において計算された確率 p を表 2 に示す．

表 1: 被験者実験による評価値の平均

取得対象	ベースライン	提案手法 1	提案手法 2
トリプル	2.48	**3.48**	3.21
リソース	2.52	2.71	**3.81**

表 2: Fisher の直接確率検定によるベースラインと各手法との比較における p 値

	提案手法 1	提案手法 2
プロパティの取得	1.0	0.0000053
リソースの取得	0.0037	0.030

提案手法１，２の評価の平均はどちらもベースラインを上回った．また検定により，プロパティの取得における提案手法２，リソースの取得における提案手法１の結果にベースラインとの有意差が認められた．これらから，PF・ICF によるプロパティの重み付けは SPARQL 検索結果のランキングを一部改善できたと言える．

5 おわりに

本研究では RDF データベース内のデータ構造に着目し，SPARQL クエリの検索結果のランキング手法を提案した．被験者実験を行い，既存の手法よりも有効なランキングを行えることを示した．今後の課題として，これらの手法を他の RDF データセットにも適用することによる手法の汎用性の検証が挙げられる．

参考文献

[1] B. Bamba and S. Soubata, “Utilizing resource importance for ranking semantic web query,” In Semantic Web and Databases, Second International Workshoo, SWDB 2004, Toronto, Canada, August 29-30, 2004, Pevised Selected Papers, pp185-198, 2004.

[2] K. Mulay and P. S. Kumar, “SPRING: Ranking the results of SPARQL queries on Linked Data”, Proceedings of the 17th International Conference on Management of Data (COMAD), Bangalore, India, 2011.

[3] L. Page, S. Brin, R. Motwani and T. Winograd, “The pagerank citation ranking: Bringing order to the web,” Technical report, Stanford University, 1998.