RDF データベースのリソースを利活用する基盤技術の開発とその応用

理学専攻・情報科学コース
一瀬　詩織

1　はじめに

近年，Web 上の共有資源として RDF 形式を採用したデータセットの公開が活発に行われ，Web 上に大きな Semantic Web データベースが形成されつつある．これに伴い，これらのデータベースから様々な目的に応じてデータを抽出する手法が必要になってきている．本研究では RDF データセットから有用なデータを抽出することを目的とし，リンク解析に基づく二つの手法を提案する．第二章では具体的に作家のリソースを対象とし，リンク解析に基づくリソース推薦手法を説明する．第三章では RDF データセットの問い合わせに用いられる SPARQL 言語を対象に，クエリ構造を利用したクエリ検索結果のランキング手法を提案する．

2　リンク解析に基づく作家推薦

2.1　概要

Wikipedia の構造データを RDF データ化した DBpedia を活用し，作家の推薦を行う．DBpedia では様々なプロパティによる作家間の関係が記述されている．これらの関係を作家間のリンクとみなしてリンク解析を行うことで，選択した作家と関連する作家をユーザへ推薦する．

2.2　手法

RDF データベースを対象としたクエリ言語である SPARQL を用い，推薦候補を抽出する．これらの候補をそれぞれ重要度，着目作家との関連の強さの指標を用いて評価し，スコアの高いものを着目作家に対しての推薦とする．

ユーザが興味を持っている作家の DBpedia 上のリソースを着目作家 w とし，関連作家 r の定義を行う．r は w とある関係によって直接リンクしているか，それぞれあるリソース a と関係を持ち，a を介して間接的にリンクしている作家であるとする．リソース間のプロパティを $prop$，$prop'$ と表すとき，r と w との間の関係は以下のいずれかのリンク構造によって表すことができる．

$$w \leftarrow (prop) \rightarrow r \tag{1}$$

$$w \leftarrow (prop) \rightarrow a \leftarrow (prop') \rightarrow r \tag{2}$$

リンク構造上の距離に基づき，今後これらの関係をそれぞれ「距離 1 の関係」「距離 2 の関係」と呼ぶ．

着目作家により関連作家の数は異なるが，一般に 1 万以上存在する．本研究では関連作家の削減のために，予備実験として作家間の重要な関係を調査するアンケートを行った．その結果「作家の影響関係」のプロパティを関連作家の抽出に用いた．

2.2.1　作家の重要度の評価

作家の重要度を評価するため，大西ら [1] による指標 Hub Score を用いる．これはハイパーリンク解析に用いられる HITS アルゴリズムや PageRank アルゴリズムを RDF データに適用した指標であり，周囲のリソースに対する対象リソースの重要度を示す．

2.2.2　作家の関連度の評価

作家 w と関連作家 r との間には式 (1)，式 (2) のような，直接または間接のリンクが複数存在する．このときリンクの数が多くまた種類が豊富であるほど，両者の間には強い関係があると考えられる．w と r の間のリンク数を作家間の関係の強さと考え，推薦を行う．

2.3　実験

2.3.1　実験仕様

2012 年の 8 月に公開された DBpedia3.8 のデータセットを用い，データセット内の作家リソース “http://dbpedia.org/resource/Haruki_Murakami” を着目作家として，作家の重要度，関連度に基づいた推薦を行った．

2.3.2　結果と考察

重要度が高い関連作家として，Hub Score の高い上位 5 名の推薦結果を表 1 に示す．また関連度が最も高い上位 2 名の推薦結果を表 2 に，推薦された Paul Auster と着目作家との影響関係を図 1 に示す．

表 1: 関連作家の Hub Score 上位 5 件 (評価対象 374 件)

順位	作家	Hub Score	Authority Score	Resource Score
1	Stephen King	1079.8	506.0	405.0
2	H. P. Lovecraft	657.2	362.0	260.0
3	Ernest Hemingway	613.6	333.0	224.0
4	Jorge Luis Borges	521.8	307.0	208.0
5	Leo Torstoy	513.1	275.0	186.0

表 2: 関係数による推薦作家　(関係数 6)

関係数	作家
6	Cagdas Cetinkaya
6	Paul Auster

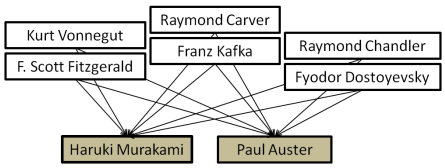

図 1: Haruki Murakami，Paul Auster 間の影響関係

表 1 では Fyodor Dostoyevsky，Stephen King 等，一般的に良く知られた作家の Hub Score が高い傾向にあることが分かる．また表 2 では Haruki と年代が近く，作風が似ているといわれる Paul Auster が推薦されており，これらの作家は目的に対して妥当な推薦が行われた結果であるといえる．

3 リンク解析に基づく SPARQL クエリ結果のランキング

3.1 概要

SPARQL は RDF データセットから条件を指定して情報抽出を行うことのできるクエリ言語である．SPARQL 言語では単純な結果のソートを行う機能を定義しているが，結果の量が多い等，結果の視認性が悪くなる場合がある．本手法では結果を重要度の順にランキングし，SPARQL クエリ検索の支援を行うことを目的とする．

3.2 手法

入力された 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，アルゴリズム内で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出力とする．

3.2.1 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 [2] を用いる．プロパティの重要度計算で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を利用した指標 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．

3.2.2 提案手法 1

検索結果から作成した，リソースをノード，プロパティをエッジとしたグラフに Bamba ら [3] 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本研究で定義した指標を用いる．

3.2.3 提案手法 2

提案手法 1 では，グラフ中のノードとエッジをそれぞれ別々に評価する．より RDF データに適した評価を行うためにノードとエッジの評価を同時に行う，トリプルの評価式を導入する．

$$TripleScore = \frac{Imp(n_s) \times PFICF(n_s, edge) \times Imp(n_o)}{linkNum(n_s) + linkNum(n_o) - 1} \quad (3)$$

n_s，n_o はそれぞれトリプルの $subjectnode$, $objectnode$ 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．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ため，ノードから出るリンクの数でスコアを分割する．

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．

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．

3.3 実験

3.3.1 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 2 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 1，提案手法 2 の 3 つの手法でランキングを行った．ベースラインには Bamba ら [3] 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た．ランキング結果の上位 20 件について，12 名の被験者が 1 〜 5 のスコアで評価を行った．

3.3.2 結果と考察

表 3: 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**

スコアの平均値を表 3 に示す．提案手法 1，2 の評価はベースラインの評価を上回り，PF・ICF によるプロパティの重み付けが SPARQL 検索結果のランキングを改善したと言える．提案手法 2 はトリプルの取得に関しては手法 1 の評価を下回ったものの，他手法よりも安定して高い評価を得られた．

4 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，データセットを用いた応用技術として DBpedia 内の作家リソースの推薦，また基盤技術としては SPARQL クエリの検索結果のランキング手法を提案した．後者では被験者実験を行い，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] K. Onishi and I. Kobayashi, “Information Enhancement on a Focused Object using Linked Data”, Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol.16, No.1. pp4-12, 2012.

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

[3] 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.