4D-4

RDF グラフ検索における検索結果の差異項抽出手法の検討

A Method for Extracting Comparable Points from Results of RDF Query Graph

飯塚 京士† 佐藤宏之† イコ・プラムディオノ† 村山 隆彦†
Kyoji Iiduka Hiroyuki Sato Iko Pramudiono Takahiko Murayama

1. はじめに

近年、Blog の普及に伴う RSS[1]や、FOAF[2]などの RDF[3]データが Internet 上に流通するようになった。リソースのメタデータを表現するためのフレームワークである RDF は、グラフ構造のデータモデルを持ち、グラフのノードに対応する各リソースは一意に示される URI で表現される。このため、同一リソースに関して異なるメタデータが記述された RDF 同士をマージし、1 つのグラフデータとして統一的に扱うことが可能となる。

RSS や FOAF に代表される各メタデータは、特定ニーズに基づいたリソース管理を目的としたボキャブラリである。例えば、RSS はニュースや記事の配信情報、FOAF は人間関係などである。これら異なるデータをマージすることで、様々な関係を表す情報を巨大なグラフ構造を持つ 1 つの RDF データとして扱うことが可能となり、個々のデータのスキーマに左右されない検索が可能になると考えている。

本稿では、RDF グラフを対象とした検索において、グラフパターンマッチングによって得られた複数の結果間の差異を抽出する手法を検討する。

2. RDF グラフ検索

RDF グラフデータから検索を行うためには、クエリに用いるグラフパターンを作成する必要がある。しかし、予めデータ構造を知らなければグラフパターンは作成できない。また、データ構造を把握しても、クエリに用いる適切なグラフパターンの選択はエンドユーザにとって困難である。この問題に対して我々は、RDF グラフデータを解析し、特徴的なグラフパターンを抽出し、そのグラフパターンを用いてクエリを自動生成する方式 CSM (Context Structure Matching Engine)を提案した[4][5]。

RDF データはラベル付き有向グラフとして表現できる。CSM におけるグラフパターンは、アークのラベルを全て固定値とし、ノードの値を全て変数とする。また、検索キーワードを代入するノード(キーノード)と検索結果の値とするノード(ターゲットノード)が含まれる。検索では、ユーザが入力した検索キーワードがキーノードに代入され、検索結果としてターゲットノードの値を返すクエリを作成して RDF グラフ検索を行う。

2.1. 問題

RDF グラフデータを検索し、パターンにマッチする結果が多数得られた場合、ユーザは得られた値を比較し、情報を選択することになる。その際、結果を比較するための周辺情報を検索する必要が生じる。例えば、検索結果として得られたノードが人を表すものであった場合、さらにその人の所属や興味などのプロパティの値を検索して差異項を比較し、絞込むことになる。

ここで問題となるのは、新たに差異項抽出用のグラフパターンを作成するコストがかかることである。比較に相応しい差異項を定めるために、RDF データのグラフを再度探索しなければならない。検索結果の内容や属性を考慮した差異項の抽出を行おうとする場合、どこまで探索範囲を広げるべきか判断が難しくなる。また、探索の結果、必ずしも値が得られるとは限らない。

3. アプローチ

そこで我々は、新たな差異項抽出用のグラフパターンを用意せずに差異項を抽出する手法を提案する。ここでは、検索で用いたグラフパターンを再利用することを考える。

3.1. 差異項抽出クエリ

差異項抽出クエリは、検索時にキーワードが代入されたグラフパターンのキーノードを変数のまま残し、今度はターゲットノードに検索の結果得られた値を代入して作成する。

例えばキーノードに“XML”が代入された以下のクエリで検索が行われた場合を考える。

```
PREFIX  dc:       <http://purl.org/dc/elements/1.1/>
PREFIX  product:  <http://context.nttlabs.com/2004/11/product#/>
SELECT  ?target
WHERE{  ?node1 <dc:subject> "XML".
        ?node1 <dc:creator> ?target.
        ?node2 <product:技術用語> "XML".
        ?node2 <product:担当者> ?target. }
```

?target で示されたターゲットノードに相当する値として<person:山田太郎>が得られたとする。差異項抽出クエリは、以下に示すように、この検索結果を上記クエリの変数であった?target 部分に代入し、今度はキーノードを変数としたものである。

```
PREFIX  dc:       <http://purl.org/dc/elements/1.1/>
PREFIX  product:  <http://context.nttlabs.com/2004/11/product#/>
PREFIX  person:   <http://context.nttlabs.com/2004/11/person#>
SELECT  ?key
WHERE{  ?node1 <dc:subject> ?key.
        ?node1 <dc:creator> <person:山田太郎>.
        ?node2 <product:技術用語> ?key.
        ?node2 <product:担当者> <person:山田太郎>.   }
```

3.2. 検索結果の差異項抽出

差異項抽出クエリを作成して追加検索を行い、新たに得られた差異項同士を比較すると、最初の検索結果の差異が明確になる。

例えば、上記最初の検索クエリで検索した結果、<person:山田太郎>と<person:田中二郎>が得られたとする。この 2 つの結果から差異項抽出クエリを生成して追加検索

† 日本電信電話株式会社 NTT 情報流通プラットフォーム研究所,
NTT Information Sharing Platform Laboratories, NTT Corporation

を行う。その結果、<person:山田太郎>には“Web サービス”、<person:田中二郎>には“RDF”という差異項が得られたとする。ここから、山田太郎さんは Web サービスで使われるプロトコル記述言語としての XML に詳しく、田中二郎さんは RDF を記述するデータ記述言語としての XML に詳しいのではないかということが言える。

4. 実験

CSM の検索結果に本手法を適用した。実験で用いたデータは、研究所内の 4 種類のオフィスデータ(論文 DB 約 3400 件、名簿 DB 約 500 件、学会活動 DB 約 300 件、プロダクト DB 約 100 件)を RDF 化し、マージしたものを使用した。実験で用いたグラフパターンは、上記 RDF データを基に CSM が自動生成したものを用いた。

4.1. 結果

キーワード“RDF”で以下のクエリを用いて検索した結果を示す。

```
PREFIX   rm:   <http://context.nttlabs.com/2004/11/researchMaterial#/>
SELECT   ?target
WHERE{   ?node1 <rm:著者キーワード> "RDF".
         ?node1 <rm:著者> ?target.
         ?node2 <rm:著者キーワード> "RDF".
         ?node2 <rm:著者> ?target.    }
```

検索の結果、26 件の検索結果(人名)が得られた。次に各検索結果に対して差異項抽出を行った結果、検索結果あたり平均 11 個、計 86 個の差異項が得られた。表 1 に、抽出した差異項の一部を示す。

表 1　抽出した差異項

差異項	差異項を持つ結果数	被覆率
セマンティック web	24	92.3%
XML	17	65.3%
メタデータ	16	61.5%
オントロジ	13	50.0%
…	…	…
概念ベース	4	15.4%
パーソナル・レポジトリ	3	11.5%
共起シソーラス	3	11.5%
…	…	…

ある差異項 *item* に対して、全検索結果に対する *item* を差異項として持つものの割合を被覆率とする。被覆率が高い差異項は、検索キーワード“RDF”と関連が深い単語が占めているのを確認した。また、被覆率が低くなるほど 検索キーワードから遠い意味をなす差異項が現れ、検索結果固有の特徴を示す傾向があることも確認した。

表 2 に検索結果“電々太郎”※の差異項の例を示す。

表 2　“電々太郎”の差異項

検索結果	差異項	重複件数	重複偏差
電々太郎	概念ベース	9 件	6.18
	パーソナル・レポジトリ	5 件	2.18
	セマンティック web	2 件	-0.81
	共起シソーラス	2 件	-0.81
	…	…	…

得られた差異項には重複があり、重複件数の多いものほど検索結果と関係が深くなる。検索結果ごとに得られた差異項に対して、重複件数の偏差を重複偏差とする。重複偏差が高いほど、検索結果と強い関わりがある差異項と言える。

※実在する人名のため、本稿では仮名(“電々太郎”)で表記する。

実験の結果、“電々太郎”に対して“概念ベース”(重複偏差 6.18,被覆率 15.4%)など、特徴的な差異項の存在を確認した。

異なるグラフパターンを用い検索で実験した結果でも差異項は抽出された。著しく特徴的な差異項は見当たらなかったが、各検索結果あたりの異項数の延べ数に差が出ていることを確認した。

5. 考察

今回の実験では、比較的単純な構造のグラフパターンを使用した。実験の結果、検索キーワードと関連深い項目や、検索結果ごとに特徴的な差異項が抽出されることを確認した。本手法で得た差異項の分布状況を解析することで、オーソドックスに“セマンティック Web”に関係するグループや、“RDF”を“概念ベース”など特徴的な用途で使っている人など、検索結果のグルーピングを行うことも可能となる。

使用するグラフパターンが複雑になると、検索に強い制約が課せられることになるため、検索件数は少なくなる傾向がある。その場合本手法を適用すると、差異項の分散が大きくなり、検索結果のグルーピングが明瞭に浮かび上がると思われる。

差異項抽出クエリは、最初の検索結果が得られているクエリのグラフパターンを利用していることから、検索結果が得られる可能性が高いと思われる。また、得られた差異項は、絞り込み検索のキーワード候補となる。そのため、続けて行う検索の際のキーワードの追加やキーワードの切替えがスムーズになる。

6. まとめ

RDF グラフ検索の結果に対して、検索クエリに使用されるグラフパターンを再利用する差異項抽出手法を提案した。本手法を、4 種類のオフィスデータを用い、CSM で自動作成したクエリでの検索で実験を行った。実験の結果、検索結果を特徴づけるキーワードが抽出されていることを確認した。

参考文献

[1] G. Beged-Dov, D. Brickey, R. Dornfest, I. Davis, L. Dodds, J. Eisenzof, D. Galbraith, R.V. Guha, K. MachLead, E. Miller, A. Swarts, E. van der Vlist, RDF SiteSummary (RSS) 1.0, http://web.resource.org/rss/1.0/spec

[2] The Friend of a Friend (foaf) project, http://www.foaf-project.org/

[3] G. Klyne, J.J. Carroll, Resource Description Framework (RDF): Concepts and Abstract Syntax, http://www.w3.org/TR/rdf-concepts/

[4] H. Sato, K. Iiduka, T. Mukaigaito, T. Murayama, Finding Similarity and Comparability from Merged Hetero Data of the Semantic Web by Using Graph Pattern Matching,WWW2005 Workshop, Activities on Semantic Web Technologies in Japan, http://www.net.intap.or.jp/INTAP/s-web/data/www2005/10_Sato2.pdf

[5] 飯塚, 佐藤, イコ, 村山, RDF データを対象としたグラフ検索におけるクエリ生成方式の検討, 人工知能学会 SIG-SWO-A502-08, 2005.