F-045

RDF グラフ検索におけるクエリ類似性判定手法

A Method for Similarity Classification of RDF Query Graph

飯塚 京士† 山本 具英† 大友 健治† 村山 隆彦†
Kyoji Iiduka Tomohide Yamamoto Kenji Otomo Takahiko Murayama

1．はじめに

RDF は、リソース間の様々な関係性を記述できるラベル付き有向グラフとして、意味を表現できるデータモデルである。RDF における全てのリソースは URI で表現されるので、異なる場所、異なる用途のデータもマージできる。また RDF は、SPARQL などのクエリ言語を用いることで複雑な検索を行える。

このような、リソース間の関係性の表現力の高さ、データマージの容易さから、近年、様々なメタデータを Web 上に RDF で公開する(Linked Data)活動が活発化している[1]。これらデータをマージして、領域を横断した新しい価値やサービスを開拓して行こうとする機運が高まっている[2]。

2．RDF グラフ検索

SPARQL などで RDF 検索を行う場合、クエリに用いるグラフパターンを、検索対象の RDF グラフから抽出する必要がある。しかし、複数のデータソースをマージした複雑なグラフ構造の RDF の場合、適切なグラフパターンの抽出は困難な作業になる。

この問題に対して我々は、RDF グラフを解析して特徴的なグラフパターンを抽出し、クエリを自動生成する技術を提案し(iMage)、実装した [3]。iMage は、検索時のキーワードになるリソース(検索キーワード)のクラスと、結果として得たいリソース(検索ターゲット)のクラスを固定し、その 2 点の間をつなぐグラフパターンから頻出なものを抽出する。抽出したグラフパターンは、SPARQL などの RDF クエリに変換し、クエリとして利用する。

2.1 問題点

iMage でクエリを自動抽出する時、グラフパターンの構造は異なるが、有識者が見ると類似する意味を表すクエリが抽出されてしまうことがある。これは、多数のデータソースをマージする場合、データスキーマを精査しきれないことなどが原因で起こる。

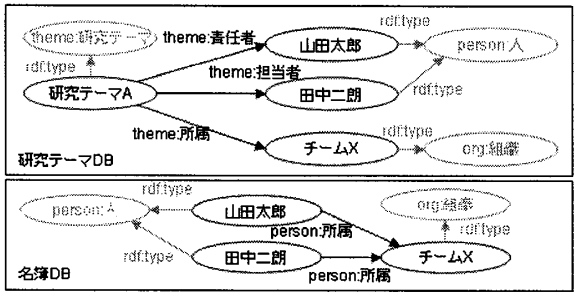

図1 2種類のRDFグラフ

以下に意味的に類似するクエリの例を挙げる。図 1 は、研究テーマ DB と名簿 DB のデータの一部を RDF グラフで表現したものである。iMage を使うと、この 2 つのデータをマージした RDF から、図 2 に示すようなグラフパターンを抽出することができる。図 2 の 3 つのグラフパターンは共に、検索キーワード“?keyword”に研究テーマを入れて検索すると、検索ターゲット“?target”に研究テーマに関係する組織が得られるクエリになる。しかし、研究テーマの所属と責任者/担当者の所属の間は強い相関関係があり、この 3 つのクエリの意味はほぼ一致する。

このような理由から、複数のデータソースをマージした RDF からは、構造は異なるが意味的に類似するクエリを抽出してしまう可能性がある。その場合、ユーザが類似クエリを取捨選択する必要が出てくる。特に、複数のデータソースをまたぐ発見的なグラフパターンを探そうとする場合、抽出されるグラフパターンの構造は複雑になり、選択候補も多量になり、ユーザの作業コストは膨大になる。

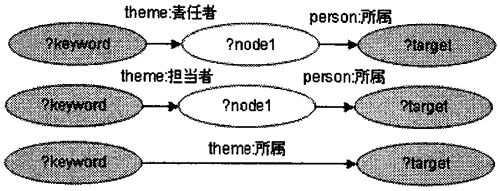

図2 自動抽出したグラフパターン

2.2 アプローチ

意味的に類似するクエリが抽出される理由として、RDF のデータスキーマの不備などがある。しかし、複数のデータソースのマージを前提とすると、データ間に隠れた依存関係の分析は、大変困難になる。

そこで我々は、データスキーマに関する検討には踏み込まない。代わりに、検索結果(検索ターゲット集合)の類似性からクエリの類似性を判定するクエリ類似性判手法を提案することで、類似クエリが多発する問題への解決を試みる。

3. クエリ類似性判定手法

大量のクエリから取捨選択作業の過程で、ユーザは次のような手順をとる。(1)クエリの種類を把握して分類し、(2)必要とする種類を選択し、(3)選択した種類の中からクエリを厳選する。この過程のうち (1) の作業は、全クエリの精査が必要な、最もコストがかかる作業である。(1)の作業に対する効果的なサポートを与えることが可能であれば、意味的な類似クエリが多発する問題に対応可能と言える。

クエリ類似性判定手法は、検索ターゲット集合の類似性からクエリの類似性を類推し、類似するクエリのクラスタを自動抽出する。この手法を用いると、ユーザによる(1)のクエリ分類作業を自動化されるため、上記問題を解決できる。

†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

3.1 クエリ類似性定義

クエリ類似性判アルゴリズムを検討するにあたり、類似性定義を定める必要がある。今回は、以下の 2 種類の定義を、厳密な類似性定義と緩い類似性定義の代表として検討を進める。

- 強類似性定義
  異なる 2 つのクエリにおいて、全ての検索キーワードで検索したとき、検索ターゲット集合が類似するならば、2 つのクエリは類似する。
- 弱類似性定義
  異なる 2 つのクエリにおいて、ある検索キーワードで検索したとき、検索ターゲット集合が類似するならば、2 つのクエリの意味は類似する。

3.2 クエリ類似判定アルゴリズム

クエリ類似性アルゴリズムを、以下のように定める。

- クエリ類似判定に用いる検索対象データは全データ。
- クエリ類似判定に用いる検索キーワードは、検索結果を持つ全ての検索キーワード。
- 検索ターゲット集合の類似判定は、検索ターゲット集合の積集合と和集合の比が閾値(類似性閾値 t_h: $0 \leqq t_h \leqq 1$)以上の場合を類似とする。

3.1 節の 2 種類の類似性定義を用いたアルゴリズムを、強類似判定アルゴリズム、弱類似判定アルゴリズムとする。

4. 実験

iMage エンジンを用いて抽出したクエリを、3.2 節で示した 2 種類のアルゴリズムを用いてクラスタ抽出を行った。

使用した RDF データは、5 種類の社内データ (論文 DB、名簿 DB、研究テーマ DB、プロダクト DB)を RDF 化してマージした約 20 万トリプルを使用した。対象とするクエリは、上記 RDF データから iMage を用いて抽出した 330 個を使用した。

4.1 結果

実験の結果、抽出したクラスタの数を表 1 に表す。

表 1 抽出したクラスタ数

類似性閾値 t_h	0	0.2	0.5	0.8	1.0
強類似判定	74(48)	225(8)	269(6)	319(3)	330(1)
弱類似判定	5(226)	37(117)	50(114)	84(77)	115(77)

※括弧内数値は、最大クラスタの要素数

強類似判定アルゴリズムではクラスタ粒度が小さく、類似性閾値 t_hを 1 に近づけると、要素数 2 以上のクラスタが抽出できなくなる。他方、弱類似判定アルゴリズムではクラスタ粒度が大きく、類似性閾値 t_hを 1 に近づけても細かなクラスタに分解せず、1 つの大きな塊が残ってしまう。

4.2 評価

今回実験で得られたクラスタが妥当なものか評価を行った。まず、今回使用した RDF の中から有識者が意味的に類似する部分グラフパターン群を抜き出す。この類似パターンで置き換え可能なクエリは、有識者が類似クエリと判断したものとみなせる。抽出したクラスタのうち要素数が 2 以上のものに対し、この有識者判定を用いて幾つの種類のクエリがあるか評価を行った。結果を表 2 に表す。

表 2 有識者判定の結果

アルゴリズム	類似性閾値 t_h	クラスタ内の種類数 平均（最大値）
強類似判定	0.2	1.24(3)
弱類似判定	1.0	1.69(7)

強類似判定アルゴリズム(類似性閾値 t_h=0.2)のクラスタは、有識者が見てほぼ同じクエリのみで構成されているが、弱類似判定アルゴリズム(類似性閾値 t_h=1.0)は、意味が異なるクエリ群が混入する確立が高い。一方、有識者による判定で類似クエリが複数のクラスタに分解されるケースが、どちらのアルゴリズムでも発生していたが、強類似判定アルゴリズムで顕著であった。

4.3 考察

2 種類のアルゴリズム共に、類似性閾値 t_hによって分類粒度の調整ができる。しかし、弱類似判定アルゴリズムでは、類似性閾値 t_hを最大にしても、要素数 77 のクラスタができてしまい、細分化には限界がある。一方、強類似判定アルゴリズムは、分類粒度は細かくなりすぎ、類似性閾値 t_hが 1 近くでは分類として使えない。

抽出クラスタの精度に関しては、強類似判定アルゴリズムは、類似性閾値 t_hが 0 に近くても有識者の評価に近い精度の結果が得られた。一方、弱類似判定アルゴリズムでは、類似性閾値 t_hを上げても充分な精度が得られなかった。しかし、有識者判定による複数クエリが少数のクラスタに収まる傾向があり、厳密性より総攬性を重視する用途には、弱類似判定アルゴリズムが適している。

従って、2 種類のアルゴリズムを併用することで、用途に合った分類粒度・精度のクラスタを抽出ができる。

5. まとめ

RDF グラフ検索におけるクエリの類似性定義を 2 種類提示し、類似クエリの自動分類手法を提案した。

4 種類のデータをマージした RDF から抽出したクエリを用いて、2 種類のアルゴリズムで実験を行った。その結果、各アルゴリズムは、用途に応じた有効性があることを確認した。

参考文献

[1] Linking Open Data, W3C SWEO Community Project, http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData/

[2] Linked Data: Principles and State of the Art, Chris Bizer, Tom Heath, Tim Berners-Lee at WWW2008, http://www.w3.org/2008/Talks/WWW2008-W3CTrack-LOD.pdf

[3] 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.