F-035

RDF グラフ検索における部分パターンの情報量に着目したクエリ判定方法

A Method for Extracting Useful Patterns from RDF Query Graph Using Amount of Information

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

1. はじめに

グラフを使い、あるインスタンスとあるインスタンスとの関係を表現する方法は従来から様々あるが、その一つにRDF[1]がある。RDF は主語・述語・目的語のトリプルでリソースの関係情報を表現するグラフ構造データの表現方法であり、異種のデータをマージして統一的に検索や構造探索を行うことが可能である。またリレーショナルデータベース等に含まれる既存の情報なども RDF によって表現することができる。これらの特徴から現在 RSS[2]や FOAF[3]など様々なデータが RDF で表現されている。また、今後もデータ間の関係を表現する方法として RDF が使われていくと思われる。

本稿では RDF グラフデータが構成するグラフ構造のクエリを用いた検索を行うために部分グラフが持つ情報量に着目し、検索データの分散値と閾値を用いて、各クエリの意味の有無の判定をする方式を考案し、効率的なクエリ抽出の可能性を示す。

2. RDF グラフ検索

我々は RDF の特徴を生かし、知識抽出源として多量の RDF グラフデータを利用し、それらをマージ、解析し、(検索用)キーワードと(被検索用)ターゲットのグラフ構造から頻出するパターンを抽出し、その頻出パターンをクエリとして利用する技術[4][5]を提案してきた。

この技術を用いると、キーワードからターゲットを検索するとき、単純に文字列マッチングによる検索結果を表示するだけでなく、キーワードとターゲットの間には何らかのグラフ構造があること、すなわちキーワードとターゲットの間に何らかの意味があることになり、それをキーワードとターゲットの間にある知識として提示できるという高度な検索が可能となる。

2.1 RDF グラフ検索に関する課題

我々は RDF グラフ検索の方式として頻出パターンをクエリとして採用している。

しかし頻出パターンをクエリとしてそのまま採用することはクエリとしての意味の有無を問わず単に採用することであり、経験上無意味なものも多く採用されることがわかっている。またクエリ数増加のためキーワードからのターゲット検索にも余計な時間がかかることもわかっている。

これらから従来の方式のクエリ作成の効率化のためには無意味なクエリを削除する必要がある。

3. 課題に対するアプローチ

クエリについて解析すると、クエリの構造全体に意味の有無があるというよりむしろクエリを構成する部分グラフに意味の有無の差があり、無意味な部分グラフを含むクエリが全体としても無意味であることがわかってきた。

2.1 節で述べた課題を解決するためにここで削除すべきクエリの構造全体ではなく、削除すべきクエリの部分グラフに着目して問題解決を図るアプローチを取ることにする。

そこで評価のために部分グラフの情報量を定義する。部分グラフの情報量とはその部分グラフの持つノード及び複数ノードの異質性(普遍性とは逆の意)の強さとする。

さらに部分グラフの情報量に着目し、部分グラフの意味の有無がその情報量で判別でき、無意味な部分グラフは情報量が少ないという仮説を立て、それによりクエリの意味の有無を判別することを考える。

まず無意味である可能性がある部分グラフを定義する。

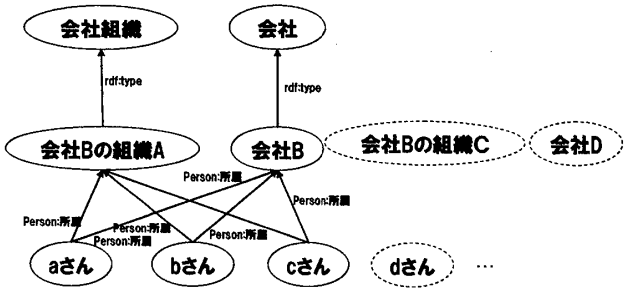

図1 ある組織・会社のRDFデータ例

人と人との関係に着目した場合、図 1 のような、所属する組織をまたいで人と人とがつながるようなグラフ構造は無意味である可能性がある。ただし後述するように、常に無意味であるわけではない。

3.1 部分グラフの分析

例えば所定の RDF データにおいて、図 1 のような部分グラフでは、一般に所属が会社である場合、所属の組織数は比較的少数であるが、それに対しそれぞれの組織に所属する人は比較的多数となる。このような構造の場合、人と人との関係で見ると多くの人が同一の会社所属という情報でつながることになり、関係性を示すデータとしてこの部分グラフが示す情報量は小さいと考える。情報量が小さいということは無意味な情報であるとする。一方、所属が会社組織のような場合、一般に所属の組織数は比較的多数であるが、それに対しそれぞれの組織に所属する人数は比較的少数となる。このような状態の場合、人と人との関係で見るとある人が所属する組織に所属する人は少数になり、関係性を示すデータとしてこの部分グラフが示す情報量は大きいと考える。情報量が大きいということは意味のある情報であるとする。

これらをそれぞれの組織に所属する人数または数学的に計算できる分散値での閾値で区別することを考える。これを図で表したものが図 2 である。

†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

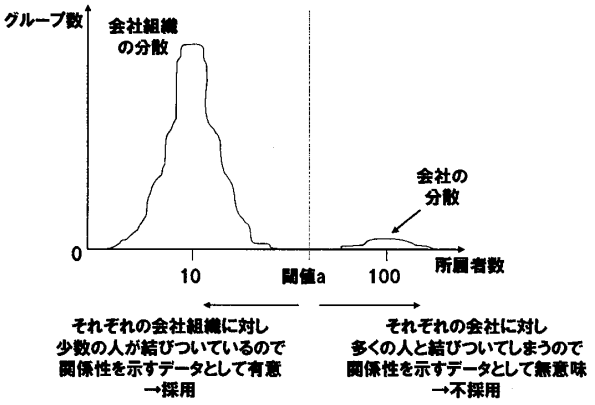

図2 出現頻度の分散表示例

図 2 で示すように図 1 の RDF プロパティの Person:所属が分かれば、所定の閾値 a より大きい所属者数を持つような部分グラフを持つクエリは無意味であり削除可能と判断でき、逆に所定の閾値 a より小さい所属者数を持つ部分グラフを持つクエリは意味があるものであり削除すべきでないと判断できる。

またこの例では所属者数で閾値を決めたが、分散値を利用する場合は情報量の小さい部分グラフをもつクエリについては分散(図 2 における会社の分散)が小さく、逆に情報量の大きい部分グラフを持つクエリについては分散(図 2 における会社組織の分散)が大きくなるため、所定の閾値より分散が小さいような部分グラフを持つクエリは無意味と判断でき、逆に所定の閾値より分散が大きいような部分グラフを持つクエリは意味があるものと判断できる。

4. 提案方式

3 章で述べたアプローチの具体的な方式を説明する。

従来技術ではグラフ構造の頻出パターンを計算した後、頻出パターンが DB に記録される。提案方式ではこの段階で DB に記録された頻出パターンから無意味である可能性のある部分グラフを含む頻出パターンを SQL および Perl の正規表現等を利用して抽出する。

抽出された無意味の可能性のある部分グラフを含む頻出パターンについて、無意味と考えられる要素を SPARQL を利用して抽出し、重複を取り除いた各要素を取得する。図 1 の例では人についてである。

また同様に図 1 の例の所属先についても SPARQL を利用して抽出し、重複を取り除いた各所属先を取る。

重複を取り除いた無意味と考えられる人を横軸、重複を取り除いた所属先を縦軸に取りプロットしたグラフが図 2 である。

理想的には重複を取り除いた無意味と考えられる要素が大きくかつ重複を取り除いた所属先が小さいような分布を示すものと、逆に重複を取り除いた無意味と考えられる要素が小さくかつ重複を取り除いた所属先が大きいような分布を示すものとの 2 種類に分類でき、前者が無意味と考えてよい部分グラフ、後者が意味を持つ部分グラフと考える。

閾値は人数による閾値を用いてもよく、また分散を計算し、その分散値による閾値を用いてもよい。

5. 考察

今回提案した方式では、まず無意味と思われる部分グラフを人間が定義しなければならないという問題がある。これは利用する RDF の性質によってもばらつきはあるもののある程度は経験的に蓄積できると考える。しかし、この部分は従来人間が行っていたことのアルゴリズム化の 1 過程であり、この部分についても今後踏み込んでいきたい。

また、分類を効果的に行う閾値を決めるという課題もある。これはデータ依存にはなるであろうが今後実データで検証し決定していきたいと考えている。

さらに、閾値による区切りで全体としては無意味と判断したものの意味のある部分グラフが残っている可能性があるという懸念がある。これについては今後実データで検証したいと考えている。

6. まとめ

本論文では我々が従来提案してきた多量の RDF グラフデータからの頻出パターンをクエリとして利用する技術に存在する課題を解決するため、部分グラフが持つ情報量に着目し、検索データの分散値と閾値を用いて、各クエリの意味の有無の判定をする方式を考案し、効率的なクエリ抽出の可能性を示した。

今後は所内の実データに対し提案方式を適用し、提案方式の仮説検証を進める予定である。また、提案仮説の課題についても自動化を目指して検討を進める予定である。

参考文献

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

[2] 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

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

[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/sweb/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.