DBSJ Japanese Journal
Vol. 17-J, Article No. 1, March 2019

RDF データに対するグラフパターンマイニング
Graph Pattern Mining for RDF Data

廣橋 美紀♡ 兼岩 憲♢

Miki HIROHASHI Ken KANEIWA

近年，セマンティック **Web** におけるリンクトデータの広がりにより，経済や化学など様々な分野においてナレッジベースとして **RDF** データはその規模を拡大している．膨大な **RDF** データを利用するためには，既存のデータから有益な知識を発見するマイニング技術が重要となっている．本研究では，多様な意味的関係を含む **RDF** のグラフ構造から述語パスと目的語の頻出パターンとルールをマイニングする方法を提案する．評価実験では，**DBpedia** と **YAGO** の実 **RDF** グラフからリソースのもつ特徴だけでなく，リソース間の意味的関係を効率的に抽出できることを示す．

In recent years, the scale of RDF data as a knowledge base in various fields, such as economics and chemistry, has been enlarged by the spread of Linked Data on the Semantic Web. Mining technologies to discover valuable knowledge from enormous RDF data become important for reusing RDF data on the Web. In this paper, we propose a method for mining frequent patterns and rules from RDF graphs that extracts various semantic relationships consisting of predicate paths and objects. Our experiments show that not only the features of resources but also the semantic relationships between resources can be effectively extracted from RDF graphs in DBpedia and YAGO.

1. はじめに

セマンティック Web の普及により，RDF（Resource Description Framework）[6, 12] で記述された膨大なリンクトデータが広く公開されている．RDF は，Web 上でリソースのグローバルな識別とリソース間の関係を記述できる機械可読な表現方法である．多様な意味的関係を表したリンクトデータは，質問応答システムやテキスト解析など様々なアプリケーションに使用されている．しかし実際にリンクトデータを利用するには，その膨大な RDF データから有益な知識をマイニングする必要がある．

RDF のような半構造化データから高頻度な属性のルールをマイニングする研究として，RDF2Rules[11] や頻出パス列挙手法 [10] がある．RDF データは，関係データベースと異なり，グラフ構造のため属性記述の自由度が高い．また，開世界仮説に基づくので，真である情報がデータベースから欠落している可能性がある．RDF2Rules では，属性間の関係性をマイニングして，RDF データ内で欠落している関係性を導く．しかし属性（述語）のみを特徴抽出するので，属性値を含んだ特徴を抽出できない．また，リソース間の関係性を導く上でループ構造に限られるので，多様な関係を扱うことができない．

本論文では，冗長で多様な意味的関係を含む RDF データのグラフ構造（RDF グラフと呼ぶ）から頻出パターンを発見するマイニング手法を提案する．特に，RDF グラフに含まれる有益な特徴パターンを見つけるために，RDF グラフの部分構造（述語パス，PRO パス，選言 PRO パス）を新たに定義する．これらの部分構造を用いて次の 2 点を実現する．

- 選言 PRO パスによる網羅的な特徴の抽出
- 述語パス，PRO パス，選言 PRO パスの組み合わせによるパターンの抽出

RDF2Rules では，目的語を除く述語のみを抽出していたが，本論文では，述語パスから目的語へ到達する PRO パスを抽出する．しかし PRO パスは，目的語（属性値）を含むので，特徴の出現頻度が少なくなり，頻出パターンとなりにくい．例えば，図 1 の RDF データにおいてファインマン，レントゲン，アインシュタイン，ローゼンのうち，3 人以上に該当するパスを頻出パターンとする．このとき，(母校, キャンパス, 都市部) のパスが見つかる．それに対し，(誕生場所, アメリカ) と (誕生場所, ドイツ) はそれぞれ 2 人しか該当しないので頻出とならない．そこで本研究では，(誕生場所, アメリカ ∪ ドイツ) といった選言 PRO パスを特徴として，合計 4 人に該当する頻出パターンを抽出する．さらに，RDF データの多様性に対して異なる視点で頻出パターンを見つけるために，選言 PRO パスだけでなく，述語パスと PRO パスを組み合わせる．

本論文の構成は以下の通りである．2 章では，本研究で扱う RDF データとマイニングの基本的な概念について述べる．3 章では，RDF データからマイニングをする際に特徴となるパスとそのパスを組み合わせたパターンの抽出方法を述べる．4 章では，マイニングにより，実 RDF データとして DBpedia と YAGO からリソースのもつ特徴だけでなく，意味的な関係を含む有益な特徴が得られることを示す．5 章では，関連研究と比較をして，本研究の有用性について述べる．最後に，6 章で結論と今度の課題を述べる．

2. 準備

2.1 RDF（Resource Description Framework）

リソースとは，Web 上の文書や情報だけでなく，Web 上にない人やものなどすべての対象物のことである．リソースを表現する際，Web 上で全てのリソースを一意に表現するために，グローバルな識別子として URI（Uniform Resource Identifier）[2] を用いる．

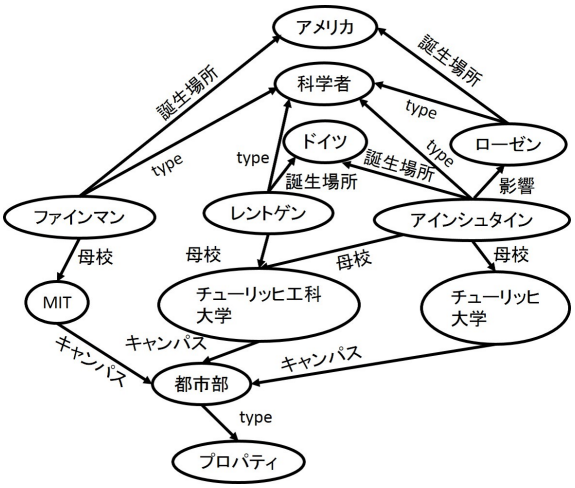

図 1: RDF グラフ例

♡ 非会員 電気通信大学大学院情報理工学研究科
hirohashi@mail.uec.jp
♢ 正会員 電気通信大学大学院情報理工学研究科
kaneiwa@uec.ac.jp

RDF は，Web 上でリソースの属性情報やリソース間の関係性を記述する枠組みである．主語（subject），述語（predicate），目的語（object） の 3 つ組を RDF トリプルと呼び，RDF データは，その集合で表される．主語と目的語は，URI が表すリソース，述語は，URI により主語と目的語の関係を示す．目的語は，URI が表すリソースに加え，リテラル（文字列）を用いてデータ値を記述できる．また，主語と目的語においては，空白ノードを用いて Web 上に存在しない，即ち，URI を割り当てられていないリソースを表現できる．URI が表すリソース集合を U，空白ノードの集合を B，リテラルの集合を L とするとき，RDF トリプル (s,p,o) の各要素は，$s \in U \cup B$，$p \in U$，$o \in U \cup B \cup L$ であり，$(s,p,o) \in (U \cup B) \times U \times (U \cup B \cup L)$ となる．RDF データ $\{(s_1,p_1,o_1),\cdots,(s_n,p_n,o_n)\}$ は，s_i と o_i をノード，p_i をエッジとする RDF グラフ G を構成する．図 1 は，科学者 4 人に関する RDF グラフの一部を示す．

2.2 RDF データに対する Apriori アルゴリズムの適用

RDF データは，グラフ構造であるため，トランザクションデータを対象とした Apriori[1] アルゴリズムを単純に適用できない．RDF データでは，述語 p と目的語 o の組 (p,o) をアイテムとすれば，その組集合をアイテムリストとしたトランザクションデータとみなせる．図 1 の RDF データは，表 1 のようにトランザクションデータ形式へ変換される．

表 1: トランザクション形式の RDF データ

人物	(p,o) の組
アインシュタイン	(type, 科学者) (影響, ローゼン) (誕生場所, ドイツ) (母校, チューリッヒ大学) (母校, チューリッヒ工科大学)
ファインマン	(type, 科学者) (誕生場所, アメリカ) (母校, MIT)
ローゼン	(type, 科学者) (誕生場所, アメリカ)
レントゲン	(type, 科学者) (誕生場所, ドイツ) (母校, チューリッヒ工科大学)

表 1 のトランザクションデータ，最小支持度 $minSup = 0.5$ を入力として Apriori アルゴリズムを適用しアイテム数 2 の頻出アイテム集合を求める．この例では，アイテム数 1 のアイテム集合を 4 個得るが，最小支持度 $minSup$ を 0.6 に上げると，(type, 科学者) の 1 つしか得られず，母校や誕生場所に関する有益なアイテムは頻出とはならない．このように最小支持度が高いと，有益なアイテム集合を得られず，反対に低く設定しすぎると，有益でないアイテム集合を含むのでマイニングに非常に多くの時間がかかる．

2.3 RDF データにおける特徴

前節では述語と目的語のみをアイテムしていたが，RDF のグラフ構造からは，述語 p と目的語 o の連なりから特徴を得ることができる．以下のように RDF グラフの部分構造であるパスを用いた 3 つの特徴表現を定義する．

定義 1 (RDF パス) RDF グラフ G が RDF トリプル (s,p_1,r_1), $(r_1,p_2,r_2),\cdots,(r_{n-1},p_n,o)$ を含むとする．このとき，RDF パスとはある主語リソース s から連なる述語と目的語であり，$(s,p_1,r_1,p_2,r_2,\ldots,r_{n-1},p_n,o)$ で表される．

RDF パスでは，パスを構成する全ての頂点と辺を特徴とする．一方，冗長性を除いた RDF の特徴として次のような PRO パスや述語パスがある．図 2 は，RDF データにおける特徴それぞれを示す．

定義 2 (PRO パス) [3] PRO（PRedicates to an Object）パスとは RDF パス $(s,p_1,r_1,p_2,r_2,\ldots,r_{n-1},p_n,o)$ から始点と内部頂点を削除したものであり，$(p_1,p_2,\ldots,p_n,o)$ で表される．

定義 3 (述語パス) [10] 述語パスとは RDF パス $(s,p_1,r_1,p_2,r_2,\ldots,r_{n-1},p_n,o)$ から全ての頂点を削除したものであり，$(p_1,p_2,\ldots,p_n)$ と表す．

定義 2 と定義 3 において，始点 s は PRO パスと述語パスのそれぞれの主語リソースという．

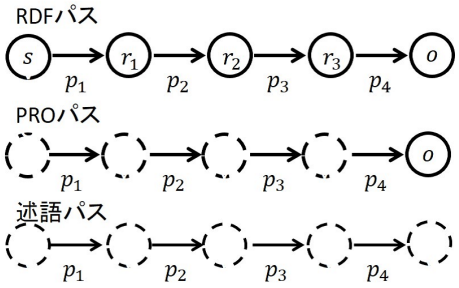

図 2: RDF データの特徴表現

3. RDF データマイニング

3.1 頻出な特徴と選言 PRO パス

RDF データにおける特徴の頻度と支持度を定義し，支持度による頻出な特徴を定義する．

定義 4 (RDF データにおける特徴の頻度と支持度) RDF データにおける対象の全主語リソース数を N，各特徴を θ，特徴 θ をもつ主語リソース集合を S_θ とする．RDF データにおける特徴の頻度を $|S_\theta|$，支持度を $|S_\theta|/N$ と定義する．

定義 5 (支持度に基づいた頻出な特徴) 下限値として最小支持度 $minSup$ を定め，支持度が $minSup$ 以上の特徴を頻出とする．特に，PRO パスの最小支持度を $minSup_{pro}$，述語パスの最小支持度を $minSup_p$ と表す．

続いて，PRO パスを一般化した選言 PRO パス [8] を導入する．PRO パスは最後に目的語が付与され，目的語を含まない述語パスに比べて具体的なので出現頻度が低くなりやすい．その結果，目的語（属性値）を含む有益なパターンが抽出されにくい．その解決策として，最小支持度を極端に下げることではなく，複数の PRO の選言を新たな特徴とする．

定義 6 (選言 PRO パス) PRO パスの最小支持度 $minSup_{pro}$ が与えられたとき，$minSup_{pro}$ 未満に設定した選言 PRO パスの最小支持度を $minSup_{dpro}$ と表す．同一の述語パス $(p_1,p_2,\ldots,p_{n-1})$ をもち $minSup_{dpro}$ 以上かつ $minSup_{pro}$ 未満の支持度である m 個 $(m \leq 5)$ の PRO パス

$$\begin{aligned}\theta_1 &= (p_1,p_2,\ldots,p_{n-1},o_1)\\ \theta_2 &= (p_1,p_2,\ldots,p_{n-1},o_2)\\ &\vdots\\ \theta_m &= (p_1,p_2,\ldots,p_{n-1},o_m)\end{aligned}$$

の主語リソース集合をそれぞれ $S_{\theta_1},S_{\theta_2},\cdots,S_{\theta_m}$ とする．

このとき $minSup_{pro} \leq \left|\bigcup S_{\theta_m}\right|/N$ を満たすならば，$(p_1, p_2, \ldots, p_{n-1}, o_1 \cup o_2 \cup \cdots \cup o_m)$ は選言 PRO パスである．尚，選言 PRO パスはわずかに出現頻度が満たない数個の PRO パスから構成することを想定しているため，多くても 5 個以下 $(m \leq 5)$ に制限している．

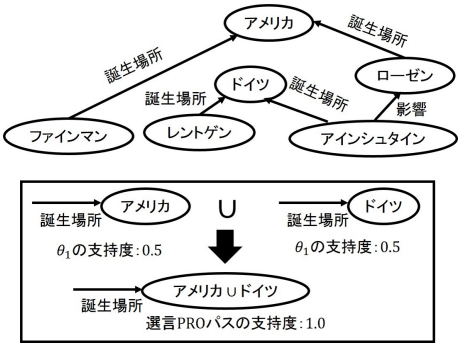

図 3: 選言 PRO パスの例

支持度が高いときに抽出されない “誕生場所” に関する特徴を選言 PRO パスにより抽出する方法を図 3 に示す．述語パス (誕生場所) をもつ PRO パスは，θ_1 = (誕生場所，アメリカ) と θ_2 = (誕生場所，ドイツ) である．しかし，いずれの特徴も支持度 2/4=0.5 のため $minSup_{pro}$ = 0.6 の場合，どちらも頻出とならない．ここで選言 PRO パスの最小支持度を $minSup_{dpro}$ = 0.5 とすると，$|S_{\theta_1} \cup S_{\theta_2}|/N$ = 1.0 は，$minSup_{pro}$ 以上の支持度となる．従って，選言 PRO パスとして (誕生場所，アメリカ ∪ ドイツ) を得る．

3.2 グラフパターンマイニングとルール生成

PRO パス，述語パス，選言 PRO パス（3 つのパスを総称して特徴パスと呼ぶ）から成る頻出のグラフパターンの抽出方法を説明する．例えば，図 4(a) は複数の特徴パスを組み合わせて表現されたグラフパターンである．これは，物理学または化学を専攻しており，母校?y が都市にあり，母校?z が公立であることを示す．図 4(b) は，グラフパターンの一部を条件部，残りの特徴パスを結論部としたルールである．物理学または，化学を専攻するならば，母校のキャンパスは都市部にあり，母校が公立大学であることを導く．

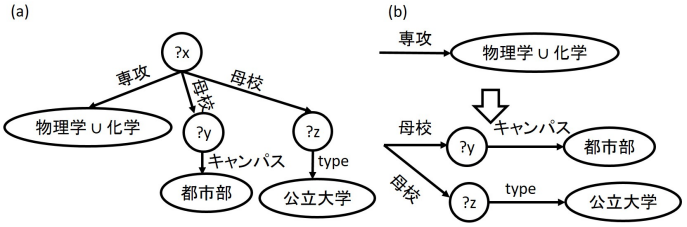

図 4: グラフパターンとルール

グラフパターンマイニングは，図 5 のように (1) クエリ解決により対象となる主語リソース集合の取得，(2) 特徴パスの探索，(3)(2) で獲得した特徴パスをアイテムとした Apriori によるカウントの手順で行う．

RDF グラフ G に対して，特徴パスの探索は，PRO パスを伸ばし，得られた PRO パスを枝刈りして行う．まず，伸ばす PRO パスを選別するために PRO パスを構成する述語パスの最小支持度を $minSup_{ex}$ と表す．図 6 のように $minSup_{ex}$ 以上の述語パスをもつ長さ k の PRO パスの集合を C_k とし，最小支持度 $minSup_p, minSup_{pro}, minSup_{dpro}$ に基づいた頻出な特徴パスの集合を L_k とする．次に L_k に対して，述語パスと目的語によって枝刈りした上で，それらの特徴パスを用いてグラフパターンを作成する．

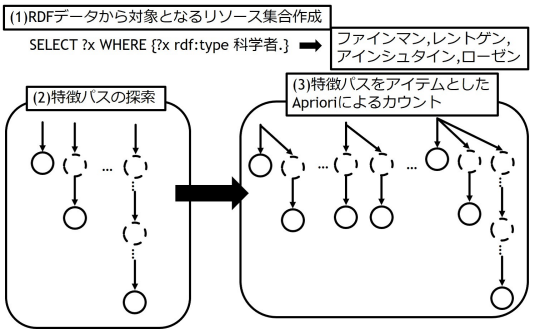

図 5: グラフパターンマイニング

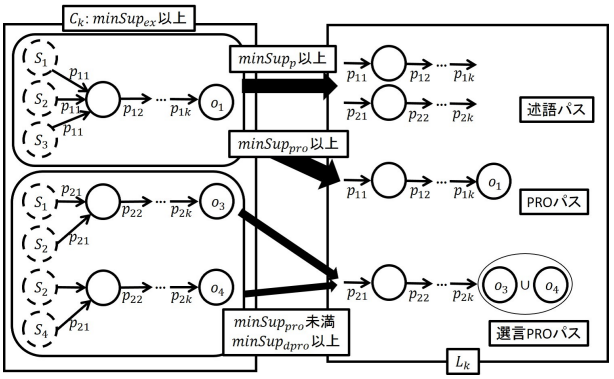

図 6: 特徴パスの抽出

3.3 特徴パスの探索

特徴パスを拡張して探索するために，図 7 は，PRO パスを長さ k から $k+1$ への伸ばす方法を示す．長い PRO パスは出現頻度が高くなり，抽出される特徴数が過剰になる．そのため特徴パス抽出と同様，伸ばす候補となる PRO パスを $minSup_{ex}$ により選別する．$minSup_{ex}$ 以上の支持度である述語パスを持つ長さ k の PRO パスの集合を C_k とする．そのうち，図 7(a) のように長さ k の述語パス $(p_1, p_2, \ldots, p_k)$ が同じで，最後の目的語のみ異なる PRO パスを探す．それらのパスの目的語 o_1 と o_2 を主語として伸びた述語 p_{k+1} と目的語 o_{k+1} を探索する．図 7(b) のように o_1 と o_2 が同じ述語 p_{k+1} と目的語 o_{k+1} をもつならば，図 7(c) のように長さ $k+1$ の PRO パスとなるので，図 7(a) の主語リソース S_1, S_2, S_3 が共通にもつ特徴パスの候補とする．

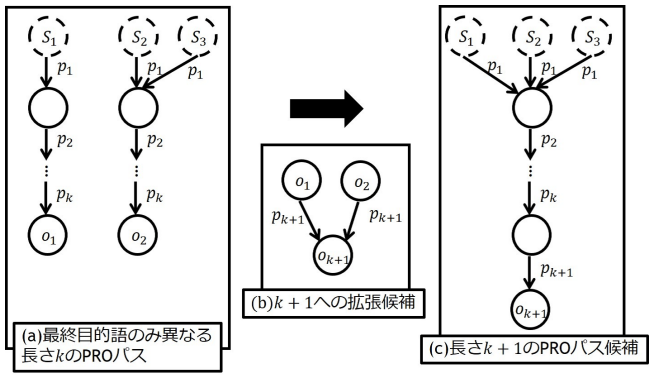

図 7: 長さ $k+1$ への PRO パスの伸ばし方

ここで，PRO パスを伸ばすとその頻度が高くなる例を示す．図 1 において $minSup_{ex}$ = 0.25，$minSup_{pro}$ = 0.6，とすると，C_1 に含まれる長さ 1 の PRO パス (母校，MIT)，(母校，チューリッヒ工科大学)，(母校，チューリッヒ大学) はいずれも $minSup_{pro}$ 以上にならない．次に，図 8 で $minSup_{ex}$ 以上の支持度である述語パス (母校) をもつ長さ 1 の PRO パスを伸ばして長さ 2 の PRO パス

を探索する．(母校) とつながる目的語は，全て同一の述語と目的語 (キャンパス，都市部) につながることから，C_2 のように長さ 2 の PRO パス (母校，キャンパス，都市部) が得られる．この結果，支持度 3/4 = 0.75 は $minSup_{pro}$ = 0.6 以上の条件を満たす．このとき，同じ主語リソースへ複数の主語がある場合に頻度の計算に注意が必要である．即ち，図 8 のアインシュタインに対して（2 つの経路による）PRO パス (母校，キャンパス，都市部) の頻度を重複してカウントしないために，この PRO パスの主語リソース集合の和集合をとる．

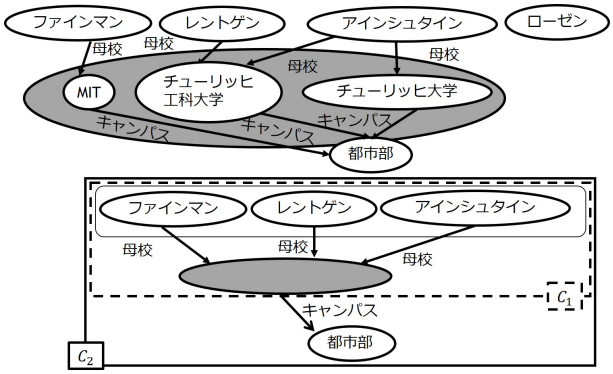

図 8: 長さ 2 の特徴パス作成例

3.4 特徴パスの枝刈り

3.4.1 述語パスの枝刈り

長い特徴パスは頻度が高くなるが，際限なく長くしても良い特徴を得られない．そこで，長さ k の述語パス $\theta = (p_1, p_2, \ldots, p_k)$ に対して，PRO パスの割合 $ratio_\theta$ を以下に定める．

$$ratio_\theta = \frac{\left| \bigcup_{\theta' \in PROPath_{\theta,k}} S_{\theta'} \right|}{|S_\theta|}$$

本研究では，$ratio_\theta \geq 0.5$ のとき，θ を長さ $k+1$ の述語パスの候補から外して枝刈りする．ここで $PROPath_{\theta,m}$ は，$minSup_{pro}$ 以上の支持度かつ述語パス θ をもつ長さ k の PRO パス θ' の集合とする．また，S_θ と $S_{\theta'}$ は，θ と θ' の主語リソース集合とする．述語パス θ に目的語を付与した PRO パス θ' が S_θ の 50% 以上の主語リソースの特徴になっていることを意味する．

図 9 では，C_2 に対して目的語の都市部から特徴パスを長くする候補として (type, プロパティ) がある．これは，どの母校にもつながり，PRO パス (母校，キャンパス，都市部) が得られた以上必要がない．即ち，多くの主語リソース “ファインマン”，“レントゲン”，“アインシュタイン” が 1 つの目的語 “都市部” へつながるとき，既に頻出な PRO パスを得ており，これ以上述語パスを長くしても冗長な PRO パスを抽出するだけである．

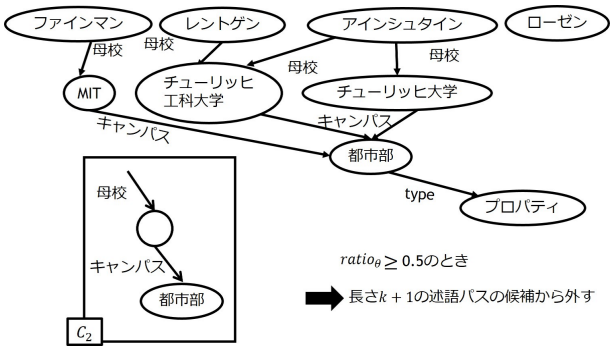

図 9: 特徴パスの枝刈り

この例では，述語パス θ =(母校，キャンパス) の主語リソース数 $|S_\theta|$ は，ファインマン，レントゲン，アインシュタインの 3 件となる．さらに，その述語パス θ から拡張可能な $minSup_{pro}$ 以上の PRO パスは，θ' = (母校，キャンパス，都市部) が該当し，その主語リソース数 $|S_{\theta'}|$ がファインマン，レントゲン，アインシュタインの 3 件となる．よって $ratio_\theta = |S_{\theta'}|/|S_\theta| = 1 \geq 0.5$ となり，(type, プロパティ) 以降の特徴を削減できる．

3.4.2 目的語による枝刈り

グラフパターンは，多くの特徴パスを組み合わせて構成される．そのとき効率良く有益な特徴パスを組み合わせるために 2 つの枝刈りを行う．

RDF グラフでは，同じ述語パスをもつ多種類の PRO パスが頻出となることがある．例えば，(誕生場所, 日本, 参加, 戦争 1), (誕生場所, 日本, 参加, 戦争 2), ⋯, (誕生場所, 日本, 参加, 戦争 k) などは誕生場所の特徴として情報過多になる．同じ述語パスの PRO パスが多いとき，1 つの述語（属性）に多数の属性値が存在して有益な特徴パスを決められない．そこで，最小支持度 $minSup_{pro}$ 以上の PRO パスを同じ述語パスに対して 5 個以内に制限する．

また，$minSup_{pro}$ 未満かつ $minSup_{dpro}$ 以上で同じ述語パスをもつ PRO パスが多く存在するとき，それらの選言 PRO パスも増大する．そこで，選言 PRO パスを構成する PRO パスを制限し，支持度が上位 5 位以内の組み合わせとする．

4. 実験

RDF ストア FROST[1][7] を用いて，本研究で提案したグラフパターンマイニングを実装し実験を行った．実験環境は OS:Windows 8.1 Enterprise 64bit,CPU: Intel(R) Xeon(R) CPU E5-2667 v2 @3.30GHz 3.30GHz, 実装メモリ 128.0GB である．各最小支持度は，$minSup_{ex} = x$，$minSup_p = 2x$，$minSup_{pro} = x$，$minSup_{dpro} = 0.3x$ とした．例えば，$minSup_{pro}$ = 0.1 のとき，$minSup_{ex}$ = 0.1，$minSup_p$ = 0.2，$minSup_{dpro}$ = 0.03 となる．また，特徴パスの最大長を 3，最大組み合わせ数を 3 と設定する．

4.1 DBpedia からのマイニング

実験対象には，DBpedia2015[2]（約 199GB）から専攻分野が物理学である科学者約 1,000 人，アメリカ企業約 7,000 件，科学者約 7 万人，映画約 20 万件のデータを用いる．また，wikilink 等の明確な意味がない述語は除外する．さらに http://dbpedia.org/property/と http://dbpedia.org/ontology/を接頭辞とする述語は内容が重複しているため，後者のみを採用する [5].

表 2 は，各対象データから得られたグラフパターン件数とその検索時間を示し，表 3 は，各対象データで得られた特徴パスの件数内訳を示している．表 2 では，$minSup_{pro}$ の値が小さくなるに従って，件数，検索時間共に線形的な増加に収まっている．最も対象とする件数が多い（約 20 万件）映画において，$minSup_{pro}$ =0.05 のグラフパターンの検索時間は，$minSup_{pro}$ =0.1 のときの約 6 倍かかる．反対に，$minSup_{pro}$ を 0.2 に引き上げると，対象データが多い科学者や映画などでは，マイニングできるグラフパターンが少なくなる．本研究では，選言 PRO パスを利用することで，表 3のように $minSup_{pro}$ が 0.05 以上の特徴を抽出できた．

表 2: グラフパターンマイニングの件数と時間（DBpedia）

	物理学者		企業		科学者		映画	
x	件数	時間 (m)	件数	時間 (m)	件数	時間 (m)	件数	時間 (m)
0.2	23,581	2.35	7,606	5.21	2	0.36	85	56.19
0.1	647,728	45.06	49,163	24.56	1,633	5.12	13,929	318.3
0.05	1,748,888	124.79	127,626	54.93	114,335	207.81	123,329	1,981.21

[1] http://www.sw.cei.uec.ac.jp/frost/index-j.html
[2] http://dbpedia.org/

表 3: 特徴パスの内訳（DBpedia）

	物理学者			企業			科学者			映画		
x	述語	PRO	dPRO	述語	PRO	dPRO	述語	PRO	dPRO	述語	PRO	dPRO
0.2	53	38	20	24	24	6	0	0	2	4	2	4
0.1	146	146	85	50	42	33	17	11	19	38	24	21
0.05	299	348	204	64	118	40	96	108	34	122	86	86

支持度，確信度がともに高いルールは，事実（例：ある企業がアメリカにあるとき，その企業がある国の首都はワシントンである）を示すのみになりえる．そこで支持度が低く，確信度を高く設定した．また，ルール $X \Rightarrow Y$ の X を条件部，Y を結論部とするとき，条件部と結論部の相関性を示す評価指標としてリフト値が用いられる．リフト値は，次のように Y の支持度に対する $X \Rightarrow Y$ の確信度の割合で表され，その値が 1.0 以上の場合に X と Y の相関性があると評価できる [4].

$$lift = \frac{X \Rightarrow Y \text{ の確信度}}{Y \text{ の支持度}}$$

物理学者，アメリカ企業，映画では，支持度 0.1 以上 0.2 以下，確信度 0.4 以上 0.6 以下，リフト値 1.0 以上とし，条件部，結論部が述語パスになるものを除いてルールを作成した．物理学者に関して，7,920 件，アメリカ企業に関して，2,793 件，映画に関しては，679 件のルールを得られた．また，科学者に関しては，支持度 0.1 以上 0.2 以下，確信度 0.6 以上 0.8 以下のルールを 220 件得ることができた．

得られたグラフパターンから作成したルール例を示す．図 10 のようにアメリカにある企業が主に位置する場所がカリフォルニア，サンフランシスコ，ニューヨーク，ロサンゼルスのいずれかであると分かる．選言 PRO パスにより，高い傾向にある特徴を結論部で示すルールを得られた．

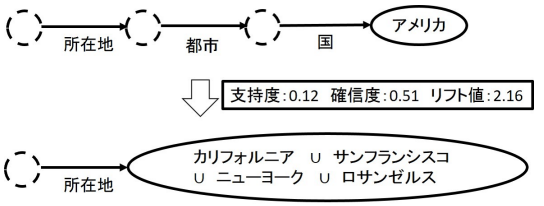

図 10: アメリカ企業に関するルール例

科学者を対象に得られたルール例は，図 11 のように出身大学がアメリカである場合，その大学は，私立大学，ランドグランド大学（理系分野に特化した大学を増やすために設立された大学），NASA のプログラム，アメリカ商務省のプログラム，州トップのいずれかに属している傾向がある．選言 PRO パスを結論部に使用することで，ある特徴をもつならば，導かれる別の特徴の候補を網羅的に示すルールを抽出できた．

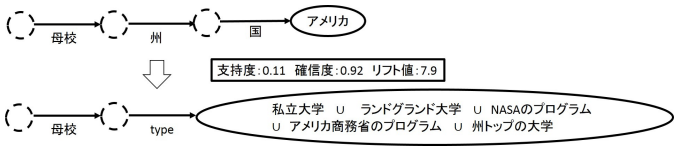

図 11: 科学者に関するルール例

4.2 YAGO の人物データからのマイニング

YAGO[3]（約 38GB）から映画に出演している女優 15,927 人，芸能人 58,616 人，画家 58,898 人，受賞経験のある人 79,245 人を対象としてマイニングを行った．表 4 は，各対象データから得られたグラフパターン件数とその検索時間を示し，表 5 は，各対象データで得られた特徴パスの件数内訳を示している．

[3]https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/

表 4: グラフパターンマイニングの件数と時間（YAGO 人物データ）

	女優		芸能人		画家		受賞者	
x	件数	時間 (m)	件数	時間 (m)	件数	時間 (m)	件数	時間 (m)
0.2	51,965	52.72	47,651	157.19	34,050	117.74	84,548	512.39
0.1	20,317	24.61	16,408	68.83	13,286	25.52	120,750	637.79
0.05	34,891	23.38	41,783	115.09	20,431	29.53	91,313	422.21

表 5: 特徴パスの内訳（YAGO 人物データ）

	女優			芸能人			画家			受賞者		
x	述語	PRO	dPRO	述語	PRO	dPRO	述語	PRO	dPRO	述語	PRO	dPRO
0.2	37	31	19	39	36	16	30	23	29	46	46	33
0.1	25	27	24	25	31	28	25	20	26	51	44	39
0.05	46	31	30	50	37	40	33	21	25	56	50	68

表 4 の受賞者のデータを除く，グラフパターン件数，検索時間，表 5 の特徴パスの内訳がいずれも $minSup_{pro}$ =0.1 のとき最小である．これは，$minSup_{pro} = 0.2$ のとき，支持度が高いため，述語パスに対する最小支持度以上の PRO パスの割合 $ratio_{\theta_k}$ が少なく，枝刈りされずに，述語パスが増えている．また，$minSup_{pro} = 0.05$ の場合は，支持度が低いため，同じ述語パスをもち目的語が異なる PRO パスが多くなり，枝刈りされ，それらの PRO パスから目的語を除いた述語パスが増えたと考えられる．

支持度 0.1 以上 0.2 以下，確信度 0.4 以上 0.6 以下，リフト値 1.0 以上とし，条件部，結論部が述語パスのみになるものを除いたとき，女優に関して 1,991 件，芸能人に関して 1,767 件，画家に関して 660 件，受賞者に関して 4,083 件のルールを得られた．

図 12 は，女優に関するデータから作成したルール例であり，ある女優が結婚しているなら，その女優が出演している映画のジャンルを示している．

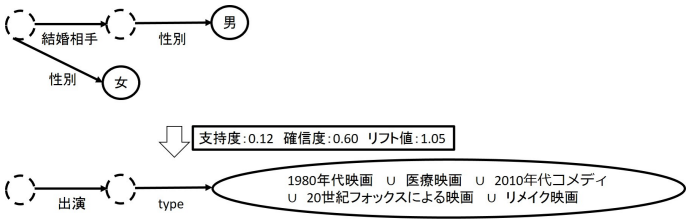

図 12: 女優に関するルール例

図 13 は，画家に関するデータにおいて，ユーロを使用する男性ならばドイツ市民であることを示す．

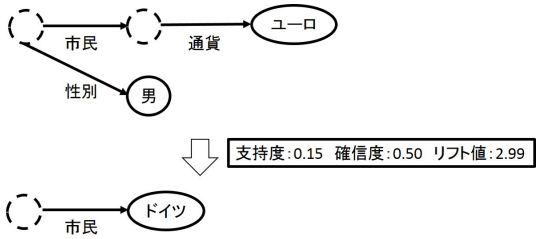

図 13: 画家に関するルール例

4.3 YAGO の事物データからのマイニング

YAGO から博物館 26,074 件，学校 45,262 件，スポーツイベント 69,144 件を対象としてマイニングを行った．表 6 は，各対象データから得られたグラフパターン件数とその検索時間を示し，表 7 は，各対象データで得られた特徴パスの件数内訳を示している．データ件数の最も多いスポーツイベントのデータを除いて，$minSup_{pro}$ =0.1 のときグラフパターン数が最大である．事物に関するデータでは，グラフパターンを構成する特徴パスの最初の述語パスがおおむね“所在地”を意味するものであった．そのため，$minSup_{pro} = 0.1$ のとき，述語パスの枝刈りが少なかったと考えられる．また，$minSup_{pro} = 0.05$ のとき，“所在地”を示す特徴パスが多く存在したため，目的語による枝刈りにより $minSup_{pro}$ =0.1 のときよりも特徴パスが削減された．

表 6: グラフパターンマイニングの件数と時間（YAGO 事物データ）

	博物館		学校		スポーツイベント	
x	件数	時間 (m)	件数	時間 (m)	件数	時間 (m)
0.2	15,367	42.17	24,050	105.45	75	0.65
0.1	79,040	180.20	55,355	228.48	7,023	64.23
0.05	47,844	76.37	17,621	73.94	35,062	88.09

表 7: 特徴パスの内訳（YAGO 事物データ）

	博物館			学校			スポーツイベント		
x	述語	PRO	dPRO	述語	PRO	dPRO	述語	PRO	dPRO
0.2	27	17	28	27	26	23	1	5	7
0.1	47	42	35	34	37	18	19	10	22
0.05	35	36	27	23	29	17	36	30	34

支持度 0.1 以上 0.3 以下，確信度 0.6 以上 1.0 以下，リフト値 1.0 以上と設定すると，博物館に関しては，49,175 件，学校に関しては，8,835 件，スポーツイベントに関しては，5,487 件のルールが得られた．YAGO の事物に関するデータからマイニングをした結果，多くが事物の存在する場所に関するルールとなっていた．これまでとは異なり，述語パスによるルールを示し，データを表現する際に使用される属性のみを示す．学校に関するデータからは，図 14 のルールが得られ，輸入している商品に関するデータが記述されている場所は，その場所の建造物とその都市がどのようなところであるかを示す．

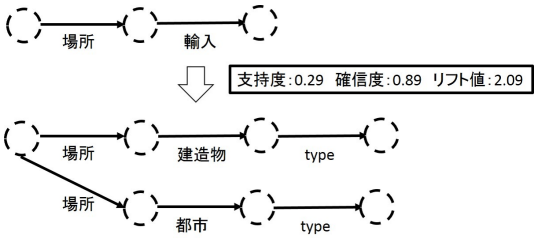

図 14: 学校に関するルール例

5. 関連研究

RDF グラフの部分構造を利用したマイニングとして，Zhichun らの RDF2Rules[11] がある．RDF2Rules では，RDF データにおいて欠落している関係性を導くルールの作成手法を提案している．頻出な述語パスをマイニングし，通常記述されるべき RDF データの関係性を見つける．しかし，述語のみを対象としており，目的語を含んだ抽出を行っていない．同様に，半構造化データを対象として，データを表現する際に使用する属性を発見する研究として頻出パス列挙手法 [10] がある．しかし，この手法は，パスの探索を開始できる箇所が根に限られており，グラフ内の複数のリソースを対象とすることが難しい．

また，その他の研究として，David らの相関ルールマイニングにおける例外ルールの抽出方法 [9] がある．例えば，相関ルールとして A ⇒ B があるとき，~ は否定を表し，A があるとき，B がないことを示す A ⇒ ~B が例外ルールの候補となる．このように相関ルールから例外ルールを作成するため，元となるルールの要素を否定したルールしか導けない．本研究では，要素の否定ではなく，支持度の低い要素を選言で合わせることで，例外として存在する要素を肯定的に表現できる．

6. おわりに

本研究では，RDF グラフに対して，述語パス，PRO パス，選言 PRO パスからなる多様な意味的関係のパターンマイニングを提案した．この方法は，冗長でノイズを含む RDF のグラフ構造から異なる特徴を組み合わせた頻出パターンの抽出を可能にしている．特に，選言 PRO パスによって目的語を含んだ特徴を逃さず発見できる．また，Apriori を応用して，大規模な RDF データからのマイニングを実現している．実際，PRO パス，述語パス，選言 PRO パスに対する最小支持度に加えて，PRO パスの拡張を選別する述語パスの最小支持度に基づいてパターン抽出の組み合わせ増大を抑えている．また，ルール作成には，支持度，確信度，リフト値を用いて，支持度を低く，確信度を高い傾向の例外的なルールを得た．選言 PRO パスを使用することで，条件や結論として複数の選択肢を示す網羅的なルールを作成できる．さらに，述語パスのルールを用いることで，不完全な RDF データの不備を発見できる．

今後の課題は，頻出パターンからの RDF データに関するルール生成の改善である．オントロジーの意味データを追加したり，複数の頻出パターン結合したりして，有用なルールを構成する方法などが考えられる．また，データの規模や種類を統計的に分析して，適切に最小支持度を決定する方法が必要である．

[謝辞]

本研究の一部は，科研費補助金（基盤研究 (C)，課題番号 18K11547）の助成を受けたものです．本稿の改善にあたり，査読者から有益なコメントをいただきました．ここに感謝いたします．

[文献]

[1] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules in large databases. In *VLDB'94, Proceedings of 20th International Conference on Very Large Data Bases, September 12-15, 1994, Santiago de Chile, Chile*, pp.487-499, 1994.

[2] Tim Berners-Lee, Roy T. Fielding, and Larry Masinter. Uniform resource identifier (URI): generic syntax. *RFC*, Vol. 3986, pp. 1–61, 2005.

[3] 荒井大地, 兼岩憲. RDF グラフの冗長な特徴表現に対するカーネル関数とその高速計算. 人工知能学会論文誌, Vol.32, No.1, pp. B-G34_1-12, 2017.

[4] Liqiang Geng and Howard J. Hamilton. Interestingness measures for data mining: A survey. *ACM Comput. Surv.*, Vol. 38, No. 3, Article No.9, 2006.

[5] 加藤文彦. DBpedia の現在: リンクトデータ・プロジェクト. 情報管理, Vol. 60, No. 5, pp. 307–315, 2017.

[6] 兼岩憲. RDF と RDF スキーマの推論. 人工知能学会論文誌, Vol.26, No.5, pp.473-481, 2011.

[7] 藤原浩司, 兼岩憲. 大規模 RDF グラフのための効率的なクエリ解決. 人工知能学会論文誌, Vol.29, No.4, pp. 364-374, 2017.

[8] 廣橋美紀, 兼岩憲. RDF データに対するグラフパターンマイニング. 第 42 回セマンティックウェブとオントロジー研究会, SIG-SWO-042-04, 2017.

[9] David Taniar, Wenny Rahayu, Vincent Lee, and Olena Daly. Exception rules in association rule mining. *Applied Mathematics and Computation*, Vol. 205, No. 2, pp. 735–750, 2008.

[10] Ke Wang and Huiqing Liu. In *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining (KDD-97), Newport Beach, California, USA*, pp. 271–274, 1997.

[11] Zhichun Wang and Juan-Zi Li. Rdf2Rules: Learning rules from RDF knowledge bases by mining frequent predicate cycles. *CoRR*, Vol.abs/1512.07734, 2015.

[12] 兼岩憲. セマンティック Web とリンクトデータ. コロナ社.

廣橋 美紀 Miki HIROHASHI

2016 年電気通信大学情報理工学部情報・通信工学科卒業．2018 年，同大学大学院情報理工学研究科博士前期課程修了．RDF データに対するマイニングの研究に従事．

兼岩 憲 Ken KANEIWA

1993～96 年富士通（株）．2001 年北陸先端科学技術大学院大学

情報科学研究科博士後期課程修了．同年国立情報学研究所情報学基礎研究系助手．2006 年 4 月～2010 年 9 月（独）情報通信機構．2010 年 10 月～2013 年 9 月岩手大学工学部電気電子・情報システム工学科准教授．2013 年 10 月より電気通信大学大学院情報理工学研究科教授．博士（情報科学）．RDF データの検索・推論・学習，オントロジーとセマンティック Web，順序ソート論理，論理プログラミング，UML ダイアグラム矛盾検証の研究に従事．情報処理学会，電子情報通信学会，人工知能学会，日本ソフトウェア科学会，IRSS（International Rough Set Society）各会員．