様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

# 科学研究費助成事業　　研究成果報告書

令和　　元 年　　８ 月 ３０ 日現在

機関番号：１０１０１
研究種目：基盤研究(C)（一般）
研究期間：2016～2018
課題番号：１６Ｋ００２９１

研究課題名（和文）極大類比 - その再構成

研究課題名（英文）Revised Algorithm for Maximal Analogies

研究代表者
原口　誠（Haraguchi, Makoto）

北海道大学・情報科学研究科・特任教授

研究者番号：４０１２８４５０

交付決定額（研究期間全体）：（直接経費）　　3,500,000円

研究成果の概要（和文）：物語や判例などにおいて，各物語に固有な登場人物などの違いを超えて観察できる類似性を検出することが本研究の目的であり，登場人物や物の間に同一もしくは類似した関係が成り立つときに二つの物語は関係構造が類似しているという．本研究では２個以上の複数の物語文を与え，一定数の物語で具体化できる共通の関係構造類似性を，行列非負分解やグラフマイニングの技法を用いて抽出できるアルゴリズムを与えた．この手法はテキスト解析のみならず，複数のデータウェアハウス間の共通関係構造抽出などにも応用でき，一般的なものである．

研究成果の学術的意義や社会的意義
構造類似性抽出タスクは，関係を構成する個体のマッチングとそれに基づく関係のマッチングという２重のマッチング問題を含んでおり，一般論として述べれば高度に組み合わせ論的な問題を内包していた．本研究では，行列非負分解によるデータ圧縮により得られるクラスタリングとクラスターを記述要素とするグラフに対するグラフマインニング技法を併せて考えることにより，現実的な計算時間で構造類似性を検出可能な道を拓いた点が学術的意義である．さらには，様々な物語のインデックスとして構造類似性を付与した物語データベースなど，応用面での基礎を与えたとの意味で社会的意義を持つであろう．

研究成果の概要（英文）：When we read several stories or legal cases, we often find relational similarities between objects in spite of the difference of stories they belong to. The goal of this research is to design and implement an efficient algorithm for finding those structural similarities among more than two domains. For this purpose, we have designed such an algorithm making use of Non-Negative Matrix Factorization and Graph Mining techniques. The algorithm we have obtained is effective not only for text analysis but also for mining similar structure among databases as data warehouses.

研究分野：データマイニング

キーワード：極大類比　疑似クリーク　局所相関推論　非負行列分解　グラフ正則化　アナロジカルヒント　クロスクラスター

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

１．研究開始当初の背景

構造類似性の意味での類比の話は古くからある．事実やイベントを個体間関係を示す述語で表現し，そうした事実からなる２つの事実集合から，個体の対応付けによりできるだけ多くの事実の対を説明できるものとして極大類比を定める研究などである．これは，個体を頂点，個体間の関係をハイパーエッジとみなした場合のグラフの部分同型(partial isomorphism) に相当する．関係の保存性という意味で，類推理論では古典的金字塔である構造写像理論に倣ったものである．結果的に，個体のマッチングと事実のマッチングという２種類のマッチング問題を解くことになり，領域記述としての事実集合のサイズにもよるが，簡単に組み合わせ爆発を生じる（ＮＰ完全問題）．その後，述語の素性構造として事実・イベントを表現できることに着目し，昔話などの比較的小規模な物語間の構造的類似性検出問題に移行した．すなわち，個体としての名詞を格（ロール）として関連づける動詞を述語とみなしたイベント列で物語を表現し，共通の部分イベント列を２つの物語の共通のプロットとして求める研究である．名詞の対応付けを分類階層で制約する，ダイクストラ法で最適コストを持つ部分列を求める等々の工夫はしたが，これも名詞の対応付けとイベント部分列マッチングの問題を扱っており，計算論的立場からは格段の改善がなされたとは言い難いものであった．さらには，主として計算効率の改善の観点から，対象文書の事前圧縮による計算コスト削減を試みた，一定の効果は確認できるが，圧縮法と圧縮結果に完全に依存している．類比とは本来多様なものであり，多様性を犠牲にしているとの意味で不満足なものであった．上記で述べた問題は，部分的な共通構造としての類比を見出すとの意味では，部分グラフ同型や関係構造マッチング問題とも異なった問題であることに注意したい．これらの問題ではグラフの内部情報を考えておらず，一方，類比の問題では，グラフの構成要素であるイベント・事実が述語の素性構造という内部データを持つからである．この内部情報を持つグラフ間の部分的な類似性・同型性を検出する研究としては，静止画像から抽出したランドマークとそれらの間の位置関係をグラフとして，２つの画像間の部分的構造類似性を最適化により求める研究が最も近い．ランドマークは属性ベクトルを内部情報として持つからである．ただし，部分的類似性の探索空間は，前述した極大類比のそれと全く同様で，対，つまり積空間であり，１００個程度の頂点の場合でも膨大な計算リソースを要するものだった．

２．研究の目的

「研究開始当初の背景」で述べたように，類比，すなわち部分的構造類似性問題においては，個体（名詞）と事実・イベント（述語の素性構造）の２重のマッチング問題を解くことになる．研究目的は算出される類比の多様性を過度に制限することなく，大幅な効率化を達成することである．さらには，事実・イベントからなる物語等の対象領域数を複数個に拡張し，所与の複数の対象領域群から一定数の領域で共通する部分的構造類似性を抽出できる手法を与え，計算論的に困難であった極大類比検出に対し現実的なアルゴリズムを開発することである．

３．研究の方法

「研究の目的」を達成するために，本研究で採用した方法論は下記の２点に集約できる：(1)「積から和へ」，および (2)「頻出パターンマイニング・グラフマイニング」の考え方と技法を取り込むことにより，少なくともスパースな対象領域に対して有効な手法を新たに設計すること．

３．研究の方法
「研究の目的」を達成するために，本研究で採用した方法論は下記の２点に集約できる：(1)「積から和へ」，および (2)「頻出パターンマイニング・グラフマイニング」の考え方と技法を取り込むことにより，少なくともスパースな対象領域に対して有効な手法を新たに設計すること．

（１）「積から和へ」

複数領域に跨る個体の対応付けは，一般には対やタプル集合となり，それだけで組み合わせ爆発が簡単に生じてしまう．これを避けるために，個体の述語における「振舞い」に着目する．すなわち，個体はどのような述語のいかなるロールとして表れるかという意味でのロール特徴を持ち，このことを個体の「述語に関する振舞い」と称する．構造類似性においては複数領域におけるイベントが対応付けられ，結果的に汎化された「抽象イベント」を得ることができる．抽象イベントにおいては個体の対・タプルは述語における共通のロールを持つことになるので，元の領域においても振舞いが同じでなければならない．これは個体の類似性・マッチングに関する必要条件であり，本研究においてはこの必要条件を満たす個体をクラスターとして纏め上げる前処理を行う．同様な振る舞いを持つ個体は異なる領域に跨って出現することから，特に「クロスクラスター」と命名する．クラスタリングは積ではなく，各対象領域に属する個体の和集合に対するものであり，この意味で「積から和」における操作となっている．

（２）　「頻出パターンマイニング・グラフマイニング」

目標とする部分的構造類似性は，その構成要素として述語の素性構造をパターン化したものである．パターンは，クロスクラスタ―を変数として持ち，少なくとも一定数以上の対象領域に対して，具体化されなければならない．これを実現可能性と言う．アイテムセットマイニングにおける頻出パターンの考え方と同じである．最も単純なパターンは，ただ一つのロールか提出確認

用らなる述語の素性構造で，これを原子素パターンと呼ぶ．一般のパターンは原子素パターンの連言である素パターンの集合として定め，原始素パターンと同様に，所与の対象領域の族に対し実現可能でなければならない．こうした実現可能性に加えて，「結束性条件」も同時にパターンに対して課す．結束性は物語の類比から派生しており，抽象イベント群において変数の共有関係において連結であることを要請する．つまり，パターンは素パターンの集まりとして連結サブグラフであることを要求する．

## ４．研究成果

（１）「積から和へ」
クロスクラスター―を得るために，述語（動詞）とそのロール（格）の対を行に，個体（名詞）を列にとったデータ行列に対する非負行列分解によって個体のクロスクラスター―を求めた．ただし，同一の個体名で異なる領域に属するものは識別するために，個体には領域ＩＤを付与する．また，構造類似性をユーザの興味にできるだけ従わせるために，類似関係にあるとユーザが判断した比較的少数の個体からなる同値関係の例示を「ヒント」として与える．本研究における行列の非負分解では，同値類を個体のグラフにおけるクリークとして表現し，各クリーク内の個体が近接する空間配置を求める．このために，グラフ正則化非負行列分解の技法を用いた．正確に述べれば，ヒントは全ての個体に対するものではないことから，グラフ正則化は一部の列に対する制約となっており，この制約を満たす部分（行）の生成系を求める際に用いる．このように，グラフ正則化非負分解と部分空間法を組み合わせた手法により，ヒントと整合した行の部分集合を得ることができる．ヒントに表れない列（個体）の相関・類似性を判定するために，部分行におけるヒント（列）が生成する列の部分空間を等価空間として定め，等化空間への射影により非ヒント属性の相関・類似性を判定する．
この方式を実装し，過失相殺に関する判例文に対し適用し，得られるクロスクラスター―の品質を評価した．クロスクラスター―の中には，ゴミ，すなわち，ユーザが与えたヒントとは意味的には離れたものも含まれるが，非負分解で得られる等化空間の斜交基底はヒントを正しく反映していることを確認した．さらには，等化空間における複数の斜交基底の合成ベクトルも射影により得られ，それらの一部はユーザの意図に照らして妥当なことも確認した．このように，パターン抽出のための事前操作としては十分なことを実証した．
（２） 「頻出パターンマイニング・グラフマイニング」
得られたクロスクラスター―を変数化し，直ちに原始素パターン集合を形成できる．原子素パターンを組み合わせて結束性条件を満たす素パターン集合としてのパターンを求める必要がある．一般には，パターンは原始素パターンの集合族となるが，集合族マイニングは集合マイニングと比較して処理が複雑になることから，原子素パターン集合を一意的に素パターン集合に翻訳できるべく，構文的な制約を課した．制約に従わないパターンは，制約に従うパターンを得た後処理としてより軽い組み合わせで求めれば良いとの考えによる
上記の構文制約により，パターンマイニング（列挙）は原始素パターン集合の列挙という最も単純な列挙問題で処理できる．一方，結束性を満たすパターンのみを生成するために，素パターン集合が 疑似クリーク，具体には k-Plex であるとのグラフ的制約も同時に課す．k-Plex は疑似クリークのモデルだと一般に理解されているが，連鎖構造からクリークまで幅広いタイプのサブグラフを求めるためにも使えることに注意すべきである．
k-Plex 列挙器の立場からのべれば，原子素パターン集合は，パターンに翻訳したものが，実現可能性と結束性という２重の制約を満たす連結サブグラフを求めるプロセスとなっており，通常の k-Plex 枚挙問題と同様の逆単調性に基づく枝刈が可能であると同時に，２重の制約により大幅な速度改善が期待できる．
研究としては，パターン列挙用に k-Plex を拡張した k-CPlex を定義し，実現可能性と結束性および連結性条件を満たす 極大 k-CPlex のための完全性を持つアルゴリズムの設計と正当性の理論的検証を完了させた．

## ５．主な発表論文等

**１．査読なし** 笹原啓佑・原口 誠： テキストデータに対する局所相関推論と構造類似性の検出，電子情報通信学会技術研究報告（言語理解とコミュニケーション）vol 118, no 439, 47--52, Feb. 2019

**２．査読なし** 原口誠・笹原啓佑：複数データ領域間共通構造マイニングの提案，人工知能学会研究会資料，SIG-FPAI-B803-15, pp. 83 - 86, Mar..2019.

**３．査読あり** Makoto Haraguchi and Yoshiaki Okubo: Finding Concepts of Music Objects with Unexpected Multi-Labels Based on Shared Subspace Method, Proceedings of the 7th International Congress on Advanced Applied Informatics - IIAI AAI'18, pp. 43 - 48, 2018.

**４．査読なし** 原口誠：データ行列上のピンポイントアナロジー，人工知能学会研究会資料, SIG-KBS-115, pp. 42 - 45, 2018.

**５．査読あり** Hongjie Zhai and Makoto Haraguchi: A Linear Algebraic Inference for Feature Association, Proceedings of the 12th Int'l Conf on Knowledge, Information and Creativity Support Systems, IEEE CPS, pp. 102 - 107, 2017.

**６．査読あり** Yoshiaki Okubo and Makoto Haraguchi: Mining Frequent Closed Set Distinguishing One Dataset from Another from a Viewpoint of Structural Index, Proceedings of the 13th International Conference on Machine Learning and Data Mining - MLDM 2017, Springer-LNAI 10358, pp. 417 - 430, 2017.

**７．査読あり** Hongjie Zhai, Makoto Haraguchi, Yoshiaki Okubo, Etsuji Tomita:
A fast and complete algorithm for enumerating pseudo-cliques in large graphs.
Int'l. J. Data Science and Analytics 2(3-4): 145-158, Springer (2016)

**〔雑誌論文〕（計　１１　件）**

**〔学会発表〕（計　２１　件）**

**〔図書〕（計　０　件）**

**〔産業財産権〕**
**○出願状況（計　０　件）**

**○取得状況（計　０　件）**

**〔その他〕特になし**

**６．研究組織**

**(1)研究分担者**
**研究分担者氏名：吉岡 真治**
**ローマ字氏名：** Yoshioka Masaharu
**所属研究機関名：北海道大学**
**部局名：大学院情報科学研究科**
**職名：教授**
**研究者番号（8 桁）：**40290879

**研究分担者氏名：大久保 好章**
**ローマ字氏名：** Okubo Yoshiaki
**所属研究機関名：北海道大学**

**部局名：大学院情報科学研究科**
**職名：助教**
**研究者番号（8 桁）：**40271639

**(2) 研究協力者　0 名**