構造類似性の列挙問題
Mining Structural Similarities

王 叡鵬 † 原口 誠 †
Wang Ruipeng M.Haraguchi

1. はじめに

構造類似性とは構造体を持つパターンによってのみ説明可能な類似性のことであり，例えばグラフの類似性も含まれるであろう．グラフの頂点は内部情報として（多重）ラベルを持つ場合の構造類似性は，グラフマイニングにおける頻出部分グラフ列挙や，グラフ間距離を用いて良く議論されている [1]．後者の距離（類似性評価指標）はグラフ全体を観たときの指標であるが，本稿では物語文等で観測可能な局所的な類似性を目標としており，多様な局所類似性の存在は必要悪だと考えている．前者のグラフパターンの場合によく使われる頻出性に基づく制御は本稿でも採用している．ただし，頂点のラベル情報には単純には還元できない素性構造を持つ「記述的パターン」によって説明可能な局所類似性を扱い，これらを列挙するための十分条件を与える．具体的にはパターン中の変数が持つ素性情報からパターンが再構成可能なことを示し，候補パターンを生成する際の無駄な組み合わせを回避する列挙器設計のための基礎とする．

2. 記述的パターンの定義

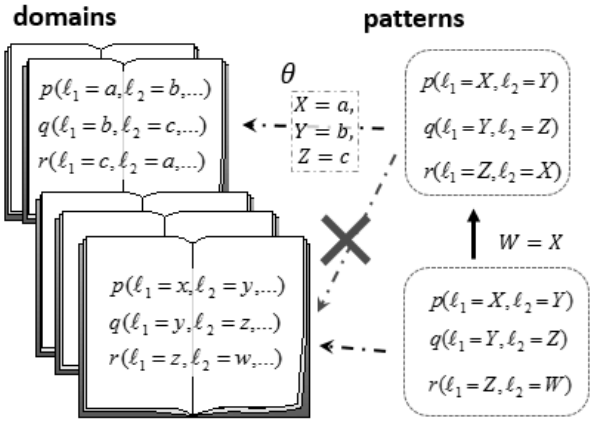
本節では記述的パターンを定義し，その検出問題を一種のマイニング問題として捉える．構造類似性は，2つの領域 (domain) 記述から，領域の要素である個体間の対応付けに基づいた関係の部分的一致として捉えるものもあったが，複数の対象領域を与えて初めて見えてくる「文脈」を考慮できてない．文脈をデータの増加により陰に把握する手法は，今日的な手法・考え方でもあり，本稿では複数領域に跨る構造類似性を論じる．

まず対象領域 D は，述語とその引数からなる基底イベント集合として与えられる．基底イベント $e = p(\ell_1 = a_1, ..., \ell_n = a_n)$ は述語 p とその素性構造 $cl = \ell_1 = a_1, ..., \ell_n = a_n$ からなり，その長さ n は固定されていない．a_j は領域中の個体で ℓ_j は識別ラベルである．テキスト処理においては，述語は動詞であり，また識別ラベルは格やロールにあたる．こうした領域 D_j のコレクション $\mathcal{D} = \{D_1, ..., D_N\}$ を入力データセットとする．特に，異なる領域は異なる個体記号を持つと仮定する．一方，述語は $\mathcal{D}$ 全体で有効であり，領域に依存しない性質や関係を記述するために用いる．

素性構造が変数で記述されるイベントを抽象イベントと呼ぶ．つまり，$e = p(\ell_1 = X_1, ..., \ell_n = X_n)$ である．パターン P とは抽象イベントの集合である．

$$P = \{..., (\ e = p(\ell_1 = X_1, ..., \ell_n = X_n)\ ), ...\}$$

†北海道大学大学院情報科学研究科
Graduate School of Information Science and Technology, Hokkaido University



パターン P は，形式的には全ての変数が存在限量化された論理式を表すが，ここでは領域をデータベース，パターンを領域に対するクエリーとして単純に考える．すなわち，P の領域 D への代入 $\theta = \{X_1 = a_1, ..., X_n = a_n\}$ で P の任意の抽象イベントが D 内のイベントとして実現できることを要請する．述語のアリティ（素性構造の長さ）は固定でないことに注意して下記の実現・支持関係 $\preceq$ を定める．

領域 D，パターン P に対し，$P \preceq D$
$\underset{def}{\Leftrightarrow}$ $\exists\theta\ \forall e = p(cl) \in P\ \exists p(cl_c) \in D$ s.t. $cl\theta \subseteq cl_c$.
ただし，cl 等は，素性リストの略記である．

例えば，$\{p(\ell_1 = a, \ell_2 = a, \ell_3 = b), q(\ell_1 = b)\}$ と $P = \{p(\ell_1 = X, \ell_2 = X)\}$ に対し，$\theta = \{X = a\}$ により $P \preceq D$ が実現される．パターン P_s, P_g 間にも全く同様に $P_s \preceq P_g$ を定める．すなわち，

$$P_g \preceq P_s \underset{def}{\Leftrightarrow}$$
$$\exists\theta\ \forall e = p(cl_g) \in P_g\ \exists p(cl_s) \in P_s \text{ s.t. } cl_g\theta \subseteq cl_s$$

$P_g \preceq P_s$ において P_s は P_g より特殊という．厳密に述べれば $P_1 \preceq P_2$ かつ $P_2 \preceq P_1$ なこともあるが，こうした「同値」なパターンに対しては同値類の代表元をとることを前提に議論を進める．

さて，所与の領域コレクション $\mathcal{D}$ に対して，本稿で欲しいパターンとは一定の個数の領域によって支持される「頻出パターン」である．

$[P] = \{D \in \mathcal{D} \mid P \preceq D\}$ P のサポート領域集合
$|[P]| \geq N\tau$ ... minsup 条件，
$0 < \tau \leq 1$ は パラメータ

パターン P を支持する $D \in \mathcal{D}$ において，D の一部のイベントや部分素性構造が捨象される．さらに個体は変数化されるので，P は $[P]$ に属する領域の共通汎化 (common generalization) であると言える．minsup 条

件は，そうした汎化がごく一部の領域に対するものであっては困るとの要請である．また，マイニングにおいて一般的に重要となる単調性（逆単調性）は本稿の枠組みでも同様に成立する．

事実１： 頻出性は逆単調な性質である．
$N\tau \leq |[P]|$ and $P_{general} \preceq P \Rightarrow\ N\tau \leq |[P_{general}]|$.
つまり，特殊パターンが頻出ならより一般的なパターンも頻出である

物語や判例文などのテキストを領域として考えるときは，イベントや名詞に対する重要度指標を上記の定義に組み込むことは可能である．例えば，パターン P が領域中に持つイメージ $P\theta$ 中に指標に照らして重要な個体が全くない場合は，そうしたパターンは minsup 条件を満たしたとしても意味がないだろう．こうした重要度指標を逆単調性を損なうことなく頻出性の議論に組み込むことは容易であるが，紙面の都合で本稿では割愛する．

記述的パターン（ＤＰ）の定義： 頻出パターンのクラスにおいて，$\preceq$ に関して極大なもの（最も特殊なもの）をＤＰと定める．

パターンの同値類を考えないときは，冗長なＤＰが多数存在する．よって同値類を考える，すなわち，既約なパターンに限定する必要があるが，紙面の都合上その議論を割愛する．

事実２： ＤＰ P は素性部分リストに関し閉じている．
すなわち，$e = p(cl) \in P$ ならば 任意の素性部分リスト $cl' \subseteq cl$ に対し，$p(cl') \in P$ である．

3. 極大頻出内包

この節では，所与の領域コレクションからＤＰを構成する際の手掛かりとなる「素性情報」について論じる．素性情報とは，パターン中の変数 X がイベントにおいてどのようなロールを担っているかの情報であり，

$$role_P(X) = \{p(\ell) \mid\ P \text{ 中のある } p(cl) \text{ で } (\ell = X) \in cl\ \}$$

として定める．ここで $p(\ell)$ は述語と素性ラベルの組を構文的に表した式で，$P = \{p_1(\ell_1 = X, \ell_2 = Y), p_2(\ell_1 = Y, \ell_2 = X)\}$ の場合は $role_P(X) = \{p_1(\ell_1), p_2(\ell_2)\}$ となる．こうした $role_P(X)$ は領域コレクションとは全く独立に，パターン P の構文のみから決まる．一方，$P \preceq D$ の場合は，代入 θ により 領域 $D \in \mathcal{D}$ が持つ個体と関連づけられる．つまり，$p(\ell = X\theta)$ が D 中のイベント $p(cl)$ の部分 （$(\ell = X\theta) \subseteq cl$）として必ず具体化でき，個体 $X\theta$ は X が P において担うロールを全て持っていることになる．このような個体 $X\theta$ と $p(\ell)$ の関係は，形式概念 [2] として扱うことができ，$X\theta$ の内包（$X\theta$ が持つ $p(\ell)$ の集合）は $role_P(X)$ を必ず含むことを意味する．詳細な議論は紙面の都合で省略するが，P がＤＰのときは，$role_P(X)$ もある形式概念の内包となり，さらに，minsup 条件を満たすことから，$role_P(X)$ の外延，すなわち，$role_P(X)$ 中の $p(\ell)$ を全て持つ個体の集合は，少なくとも $N\tau$ 個以上の領域に跨って存在することを意味する．結局，形式概念の内包にパターンと全く同じ minsup 条件を課したとき，$role_P(X)$ が持つ性質を下記の事実としてまとめることができる．

事実３： DP P に対し，$role_P(X)$ は minsup 条件を満たす（集合の包含関係に関して）極大な頻出内包である．

4. 極大頻出内包からのＤＰの再構成

パターン $P = \{..., p(..., \ell = X, ...), ...\}$ がDP の場合は，事実２から $role_P(X)$ は素性リストを分解して得られる「原始述語形」$p(\ell = X)$ に対応する $p(\ell)$ を全て含んでいる．また，事実３により，対応する $p(\ell)$ の集合は形式概念分析によりターゲットとする DP の $role_P(X)$ は必ず極大頻出内包 として列挙できる．よって，P が持つ原始述語形 $p(\ell = X)$ は $role_P(X)$ の $p(\ell)$ から復元できることは明らかだろう．P に現れる他の変数 Y に対しても，その原子述語形 $p(\ell' = Y)$ も $role_P(Y)$ から得られるので，同一の述語に対し原子述語形を合成することにより，より特殊な $p(\ell = X, \ell' = Y)$ が得られる．この操作（述語内素性形成）を繰り返せば，もとのＤＰ P は，パターンとは独立に定義される形式概念における極大頻出閉包だけを使って必ず再構成できる．述語内素性形成操作自体も minsup 条件に基づいた枝刈が可能だが，ＤＰ の抽象イベントは minsup 条件を満たすことから，枝刈されることは決してないことにも注意したい．つまり，安全な枝刈が可能であり，全ての組み合わせを試みる必要は全くない．

5. サマリーと課題

以上の考察をまとめると，DP の多様性の要因は２点に集約できる．まずは，述語内素性形成のための部品たる極大頻出閉包の数，および，述語内素性形成操作の組み合わせの可能数によりＤＰ列挙の複雑さは支配される．その複雑さが実際問題としてどの程度のものかは，領域コレクションの作り方にも依存し，例えば少数の「雛形」を組み合わせてできる場合は，比較的に容易だと推察できよう．一方，そもそも多様な類似性を秘めた入力に対しては，閉包の総数を抑える効果を持つ個体の重要度指標がすぐにでも利用でき，また，述語内形成操作に関しても，確率的に起こりやすい組み合わせを優先する標準的手法が今回の問題にも適用できると考えている．

参考文献

[1] C.C.Aggarwal: Data Mining - The Textbook, Springer (2015).

[2] B.Ganter, G.Stumme, and R.Wille (Eds): Formal Concept Analysis – Foundations and Applications, LNAI-3626, 348 pages, Springer (2005).