

1 はじめに

データマイニングとは、大量のデータから隠された知識や、新しい規則を発見するプロセスである。その中でも、関係型データマイニング(以下 MRDM)は述語論理で表現されており、比較的ユーザにとっても理解しやすい形式となっている。しかし、大規模データベースを扱う場合には、いかに論理式で表されていようがユーザにとって理解しにくいいため、理解しやすい出力形式が自然と必要となってくる。既存の出力形式としては、当研究室吉野らによって提案された視覚化表現 [1] があるが、生成されたルールには、それでも表現できない構造や、ルールごとに違った意味を持つ構造が存在する。例えば、人物同士のつながりを表す接続関係や、全体がどの部分を含むかを表す包含関係である。本研究では、これらの意味を考慮した視覚化表現を提案する。

2 既存手法の問題点

例えば、以下のようなパターン p_1, p_2 があるとする。

$$p_1 : \text{grandfather}(\text{koji}) \leftarrow \text{parent}(\text{koji}, \text{yozo}) \wedge \text{parent}(\text{yozo}, \text{koyochi}) \wedge \text{male}(\text{koyochi})$$

$$p_2 : \text{train}(t) \leftarrow \text{has_car}(t, c) \wedge \text{has_load}(c, l) \wedge \text{triangle}(l) \wedge \text{load_num}(l, 2)$$

p_1, p_2 を視覚化表現に変換した際、理想的な表現は図 1, 2 となるが、パターン種類の膨大さや、機械的に処理することを考えると現実的ではない。そこで、吉野らによって図 3 のような有向グラフでの表現が提案された。しかし、吉野らによる手法は有向グラフの頂点と辺が元々の述語の項と述語名から成る単純なものであり、視覚化表現を用いても構造が理解しにくい。さらに、包含関係を表す視覚化表現 (p_2) が表現できないという欠点が存在する。

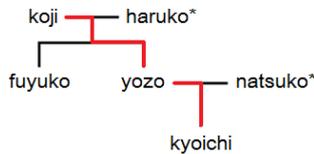


図 1: p_1 の視覚化

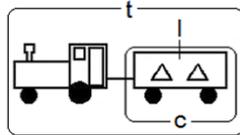


図 2: p_2 の視覚化



図 3: 吉野らによる手法

3 提案手法

パターンは以下のような 3 つの対象から構成されていると想定し、それらを分類し、利用することによりパターンを視覚化表現に変換することを考える。

- 個体 ... 具体的な対象。
- 属性 ... 個体を修飾する抽象的な対象 (個体の持つ性質)。
- 関係 ... 個体同士を関連付ける対象。

ここで、一つの項からもう一つの項を関数的に呼び出す働きをするモードを利用して、述語の述語名と項から上記 3 つの構成条件を以下のように与える (\oplus は入力引数, \ominus は出力引数, \odot は定数引数を表す)。

- 個体 ... 入力引数, 出力引数に相当。
- 属性 ... 定数引数, 出力引数のみを持つ述語の述語名, 定数引数を持つ述語の述語名に相当。
- ただし、前者 2 つを特に属性頂点、後者を特に属性辺とする。
- 関係 ... 出力引数を持つ述語の述語名に相当。

$\text{has_car}(\oplus A, \ominus B)$	$\text{has_load}(\oplus A, \ominus B)$
列車 A は貨車 B を持つ 個体 : A, B 関係 : has_car	貨車 A は貨物 B を持つ 個体 : A, B 関係 : has_load
$\text{load_num}(\oplus A, \odot B)$	$\text{triangle}(\oplus A)$
貨物 A は B 個である 個体 : A 属性 : B (頂点) load_num (辺)	貨物 A は三角形である 個体 : A 属性 : triangle (頂点)

表 1: モードから成る対象の例

上記の条件のもと p_2 について調べてみると、表 1 のようになる。

- グラフの考えをもとに、個体と属性頂点を頂点、属性辺と関係を辺だと考え、接続関係を表す連結表現と、包含関係を表す領域表現を以下のように表現する。
- 連結表現 ... 頂点と辺をつないでグラフのように表現する。
- 領域表現 ... 辺を、包含を表す囲いだと考えて、頂点の中に頂点が含まれるように表現する。

4 実行例

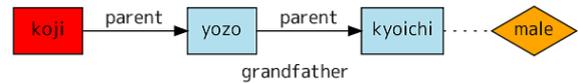


図 4: 連結表現

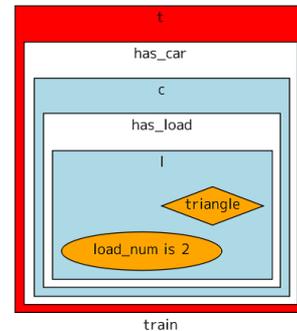


図 5: 領域表現

当研究室のマイニングシステムに組み込んで、以上のようにマイニングしたパターンを機械的に連結表現や領域表現に変換することができる (適宜頂点や辺の形や色を変えて理解しやすい様にしている)。

5 まとめと今後の課題

本研究では、MRDM におけるパターンの視覚化において、パターンを意味的に個体、属性、関係という 3 つの対象に分類できることを想定することにより、接続関係や包含関係を表す連結表現や領域表現での視覚化が可能となった。今後の課題としては、個体同士の領域を共有する場合や、接続関係と包含関係を同時に含むパターン、接続関係と包含関係以外の関係を含むパターンの視覚化表現の考察が挙げられる。

参考文献

[1] 吉野 伶, “関係型データマイニングのための統合的環境の構築”, 名古屋工業大学平成 20 年度卒業論文, 2008.