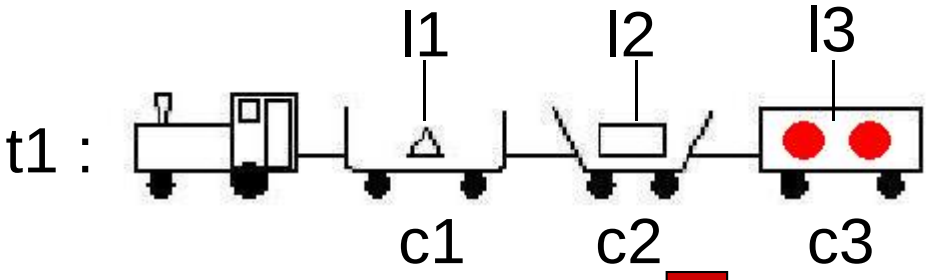
関係データベースシステムを結合した関係型データマイニング法実装の改善

名古屋工業大学　情報工学科
所属　犬塚研究室
平成17年度入学
17115128　日比野仁志

研究背景

- データマイニング
  - 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
- 関係型データマイニング
  - 一階述語論理に基づく知識発見の枠組み
    - 構造的なデータを扱うことができる
    - 論理的な記述により高い可読性をもつ

関係型データマイニング



部分構造をたどる述語

属性を記述する述語

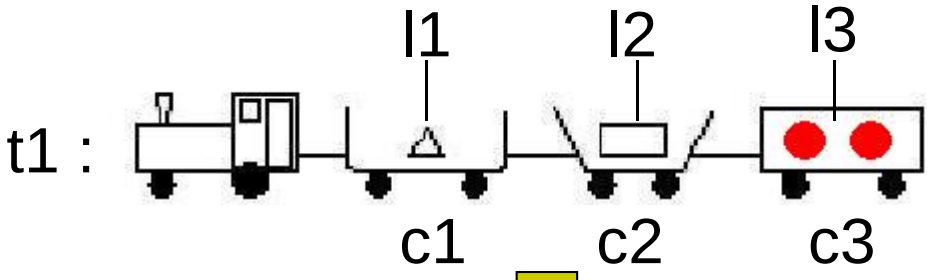
pattern : has_car(t1,c1) ∧ has_load(c1,l1) ∧ triangle(l1)
pattern : has_car(t1,c2) ∧ has_load(c2,l2) ∧ rectangle(l2)
pattern : has_car(t1,c3) ∧ has_load(c3,l3) ∧ circle(l3)
pattern : has_car(t1,c3) ∧ has_load(c3,l3) ∧ red(l3)

これらのパターンで頻度の高いものを枚挙する

MAPIXとSQL_MAPIX

- MAPIX [Motoyama ’06]
  - 性質という意味のある特徴を表す基本パターンに限定してマイニングをおこなう手法
  - マイニングの流れ
  1. 選択した事例から関連リテラル集合というものを生成
  2. 関連リテラル集合から性質を抽出
  3. 抽出した性質の組合わせから頻度の高いものを枚挙する
- SQL_MAPIX [牧野 ’07]
  - データベース上のデータをマイニング可能にした

関連リテラル集合 (MAPIX)



関連リテラル集合 = この対象についての全ての事実

has_car(t1,c1), has_car(t1,c2), has_car(t1,c3),
has_load(c1,l1), has_load(c2,l2), has_load(c3,l3),
triangle(l1), rectangle(l2), circle(l3), red(l3)

has_car(t1,c2) ∧ has_load(c2,l2) ∧ rectangle(l2)

関連リテラル集合(SQL_MAPIX)

事例のテーブル

train
t1
t2

背景知識のテーブル

east
t1

west
t2

has_car	
t1	c1
t1	c2
t2	c3

train	east	west	has_car	
t1	t1		t1	c1
t1	t1		t1	c2
t2		t2	t2	c3

t1に関する
関連リテラル集合

研究目的

- 関連リテラル集合の計算において
  テーブルサイズが膨大になる

- 背景知識のテーブルを圧縮しテーブル数を減らす方法を提案する

関連リテラル集合生成の際の問題点

事例のテーブル

train
t1
t2

テーブル: rectangle

train	load
t1	r1
t1	r2
t2	r3
t2	r4
t2	r5

テーブル: circle

train	load
t1	c1
t1	c2
t2	c3
t2	c4
t2	c5

元の行数の掛け算で決まる

train	load1	load2	
t1	r1	c1	4行 (2x2)
t1	r1	c2	
t1	r2	c1	
t1	r2	c2	
t2	r3	c3	9行 (3x3)
t2	r3	c4	
t2	r3	c5	
t2	r4	c3	
⋮	⋮	⋮	
t2	r5	c5	

解決手法

1. 述語のモードと型の一致する述語を
   グループ分けする
2. グループ分けされた述語のテーブルを
   UNIONする
3. UNIONしたテーブルを用いてSQL_MAPIXと
   同様に関連リテラル集合を作成

解決手法の例

事例のテーブル

train
t1
t2

テーブル: rectangle

train	load
t1	r1
t1	r2
t2	r3
t2	r4
t2	r5

テーブル: circle

train	load
t1	c1
t1	c2
t2	c3
t2	c4
t2	c5

元のテーブル
表す列を加

train	predicate	train	load	load
t1	rectangle	t1	r1	
t1	rectangle	t1	r2	4行
t1	rectangle	t2	c1	(2+2)
t1	rectangle	t2	c2	
t2	rectangle	t2	r3	
t2	rectangle	t1	r4	
t2	rectangle	t1	r5	6行
t2	circle	t2	c3	(3+3)
t2	circle	t2	c4	
t2	circle	t2	c5	

実験

- 3369個の事例をもつ英文の構造データ
  - ペンシルバニア大学のPenn Treebank Projectによって
    タグ付けされた構文木を基に[浦澤 ’05]が準備したもの
- 本手法により背景知識のテーブル53個を
  3個に減らした
- 実験結果
  - 従来のSQL_MAPIXでは動作しなかったが
    今回の手法により動作することが確認できた

事例数	1	10	100	1000	2000	3369
性質の数	16.8	139.4	856.3	4396.8	7428.0	10422.2

まとめと今後の課題

- まとめ
  - SQL_MAPIXを改善する手法を提案した
    - 背景知識がもつバイアス情報を利用して背景知識のテーブルを圧縮し関連リテラル集合を生成
  - 従来のSQL_MAPIXでマイニングできなかったデータをマイニングできるようになった
- 今後の課題
  - 現段階では関連リテラルのテーブルの列数が制限されているため、この制限を除去する