平成 18 年度　卒業論文概要

入学年度　平成 14 年度	学生番号　14217660	氏名　吉川 知孝
卒業研究題目 ユーザレベルを考慮した関係ルールマイニング手法の拡張		和田・犬塚　**研究室**

1. はじめに

データマイニングとは，蓄積された大量のデータから価値のある知識を取り出すプロセスである．データマイニングの手法の一つとして，事例から性質を抽出してこれをマイニングに用いる，**MAPIX** (Mining Algorithm by Property Item eXtraction) アルゴリズムが開発された．しかし，MAPIX はデータベースに対してビューを付け加えることができない，あるいは元々あるデータにはユーザの理解に及ばない部分があるなど，物理レベルデータとユーザレベルデータが一致しないという問題点があった．そこで，本研究では物理レベルとユーザレベルを明確に分け，ユーザレベルでマイニングを行う手法を提案する．

2. MAPIX

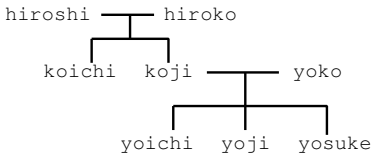

図1．家族関係

簡単な例として図1の家族関係について考える．これは，親子の関係を表す parent と性別を表す male，female という述語で構成されている．図1において hiroshi に着目したとき，hiroshi には koji という息子がいること，hiroshi には yoji という男の孫がいることは次のように表される．

parent(hiroshi，koji) ∧ male(koji).

parent(hiroshi，koji) ∧ parent(koji，yoji) ∧ male(yoji).

このように対象 hiroshi がもつ事実を表す述語の組を，hiroshi の**性質**と考えることができる．MAPIX では，性質を取り出すために**飽和節**を生成する．これは，注目した事例に関連するデータを集めて来たものであり，全ての性質が含まれている．

MAPIX の概要を次に与える

1. 与えられた事例の集合からいくつかの事例を選択し，その飽和節を生成する．
2. 飽和節の前提部のリテラルの集合から，事例に関する性質を取り出す．
3. 事例の性質を使って，興味深いパターンを枚挙する．

3. 提案手法

3.1 提案手法の概要

MAPIX は，与えられたデータベースから興味深いパターンを導出する手法である．しかし，この与えられたデータベースとは，**外延的データベース** (事実の集合としての関係の集まり) のことであり，**内包的データベース** (外延的データベースの上に定義されたルール) で定まる関係である**ビュー**を自由に追加し，ビューからパターンを導出することはできないという問題があった．また，膨大なデータを持つ外延的データベースにはユーザの理解に及ばない部分があるなど，物理レベルデータとユーザレベルデータが一致しないことが多かった。そこで、本研究では物理レベルとユーザレベルを明確に分けユーザレベルでマイニングを行う手法を提案する。

ここで，本研究で使う用語について述べる．**物理レベルデータベース**とは，外延的データベースのことである．**拡張データベース**とは，物理レベルデータベースに内包的データによって定まる全てのビューを合わせたデータベースのことである．**ユーザレベルデータベース**とは，拡張データベースのうち，ユーザの関心のあるデータでユーザ自身が宣言する．

3.2 飽和節生成アルゴリズムの改良

関係データベースからマイニングをおこなう MAPIX を，**拡張データベースからマイニングをおこなえるように拡張**する．具体的には，ビューのリテラルも飽和節の前提部のリテラルに追加することで，拡張データベースから性質を取り出すことができるようになる．これにより，データベースにビューを追加することが可能となる．

3.3 ユーザレベルデータベースに含まれる関係のみを利用したマイニング手法

拡張データベースにおいて，ユーザレベルデータベースに含まれる関係を宣言する方法について検討する．MAPIX では，性質は飽和節を用いて生成していた．飽和節を生成するアルゴリズムを改良することで，**ユーザレベルの述語のみを持つ性質を抽出**する．提案したアルゴリズムを2つ示す．

一つ目のアルゴリズムでは，飽和節を生成した後に，飽和節からユーザレベルデータベースに含まれないリテラルを削除し，これから性質を抽出する．

二つ目のアルゴリズムでは，ユーザが宣言したリテラルだけを用いて飽和節を生成し，これから性質を抽出する．

一つ目のアルゴリズムを採用したときと，二つ目のアルゴリズムを採用したときとでは，得られる飽和節は異なる．しかし，それぞれの飽和節から得られる性質は同じであることを示した．

4. 実験

4.1 飽和節を生成するのにかかる実行時間の比較

既存手法である関係データベースから飽和節を生成するアルゴリズムと，提案手法である拡張データベースから飽和節を生成するアルゴリズムを2つのデータを用いて実行時間を比較した．データベースには，家族関係と物質の突然変異性に関するデータを使用した．家族関係のデータベースには2つ，物質の突然変異性に関するデータベースには23のビューを追加した．ルールが使えない既存手法ではルールを事実に展開し，それを背景知識とした．

実験した結果，提案手法の実行時間は既存手法と同程度だということがわかった．また，ビューを用いることにより，背景知識のデータベースのデータ量が削減できたことも確認できた．

4.2 ユーザレベルデータベースからマイニングしたときの実行時間の比較

拡張データベースから飽和節を生成してマイニングする手法と，ユーザレベルデータベースに含まれる関係のみを用いて飽和節を生成してマイニングする手法との MAPIX の実行時間の比較をした．

拡張データベースからマイニングした場合，閾値が低い場合は現実的な時間で結果が出力されなかった．それに対し，ユーザレベルデータベースからマイニングした場合は閾値が低い場合でも結果を得ることができた．ユーザレベルデータベースからマイニングすることにより，大幅に実行時間が短縮されたことが確認できた．

5. まとめ

本研究では，飽和節生成アルゴリズムを改良することで，物理レベルデータベースからマイニングする MAPIX をユーザレベルデータベースからマイニングできるように拡張した．この拡張により，ビューが追加できるようになった．また，ユーザレベルデータベースに含まれる関係を宣言することで，ユーザの利便性が向上し，MAPIX の実行時間が短縮できたことを実験で示した．

参考文献

[1]Jun-ichi Motoyama, Shinpei Urazawa, Tomofumi Nakano, Nobuhiro Inuzuka : “A mining algorithm using property items extracted from sampled examples”, ILP2006(出版予定)