関係型データマイニングによる
Web 利用データマイニングのためのデータ表現

犬塚研究室　　知能系
No. 17115143　　松本 潤

1　はじめに

Web アクセスログに注目したデータマイニングを Web 利用マイニングと呼ばれ，Web ユーザの振る舞いを理解することを目的とする．一人のユーザのアクセスはセッションなどの単位で扱われ，その構造的特徴をマイニングすることは興味深い．

本研究ではアクセスの構造的特徴をマイニングするため，関係型データマイニング手法を利用する際，ログの表現の形式について検討する．

データマイニングは，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り，関係型データマイニングは，述語論理を用いたプログラムの形式で行うデータマイニングである．述語論理をもちいることにより豊かな表現空間を持ち，可読性の高い解析をおこなえると考えられている．

2　アクセスログの構造のデータ表現

関係型データマイニング法 MAPIX は，構造のあるとその構造的特徴を抽出してマイニングするためアクセスログにふさわしいと考えられる．

MAPIX では，構造を持つ対象を目的述語とし，その部分構造へアクセスする経路述語と，部分構造の属性を記述する判定述語を用いてデータを表現する．

アクセスログをセッションに区切り，セッションを構成する１つ１つのアクセスログを部分構造と捉えた．１つのログをレコードと呼び．ログレコードからそれに連なるログレコードを得る述語を基本の経路述語とし，これをいくつかの側面で部類して多様なパターンを表現できるようにした．

表１に代表的な述語を示す．判定述語は各ログの属性とし表１にその代表的なものを示す．

3　データ表現

本研究では，述語による知識表現の違いでマイニングにどのような違いがあらわれるかを検証した．まず３つの違う表現方法を用意した．

述語セット１　ログレコード間の繋がりに注目
述語セット２　ログレコードの時間差に注目
述語セット３　繋がりと時間差を同時に注目

述語セット２では，述語セット１で表現した経路述語をこまかくわけることでマイニングにどう関わるかを検証する．述語セット３では，同じ経路を２つの表現で表すことによりどのような影響を及ぼすかを検証した．

述語名 (モードとタイプ)	述語	意味
link_to (+log, -log)	経路	ログレコード同士の繋がりをしめす +log が参照元のログレコードが入り， -log にリクエストのログレコードが入る
initial_Log (+log)	目的	一回の閲覧の動作で閲覧し 始めたときに生じるログレコード
terminal_Log (+log)	判定	ログレコードから他のログレコードへ 移動していないログレコード
branch_Log (+log)	判定	複数のログレコードから 参照されているログレコード
small_gap (+log, -log)	経路	link_to で 時間の差が小さいもの
middle_gap (+log, -log)	経路	link_to で 時間の差が 平均の時間差のもの
big_gap (+log, -log)	経路	link_to で 時間の差が大きいもの

表 1: 本研究でログの構造表現に用いた述語

4　実験

上記の３つの述語表現について，実行時間とマイニングした頻出アイテムセットの個数について調べた．実験には，プロキシサーバに蓄積された本研究室の学生らの Web アクセスログをもちいている．

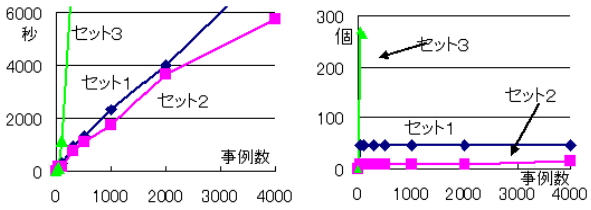

図 1: 事例に対する計算時間 (左) と頻出アイテムセット数 (右)

5　考察

図１の実験から述語セット２のように経路を複数にわけることは，時間的には変化がないが，頻出アイテムセットの個数は減少する．また述語セット３のように同じ経路を多種類で表現し同時にマイニングをおこなうとパターン数が膨大に増え実行時間もアイテムセットの個数も膨大に増加する．また図１から事項時間は，事例数に対し線形的に増加した．

6　まとめ・今後の課題

本研究では，関係型データマイニングと Web 利用マイニングに適用する際のデータ表現について検証した．関係型データマイニングでは，構造的な情報の解析にむいているため，Web 利用マイニングに適用できる．背景知識の組合せによって時間，パターンに大きく影響を与える．これらの結果を利用した背景知識の設計方法については今後の課題である．

参考文献

[1] J. Motoyama, S. Urazawa, T. Nakano and N. Inuzuka: “A Mining Algorithm Using Property Items Extracted from Sampled Examples”, ILP-2006, pp335-350, 2007.