No.87

飽和性と極大性を用いた単一系列データにおける頻出系列の高速マイニング

村田拓也 , 岩沼宏治

山梨大学医学工学総合教育部 山梨大学医学工学総合教育部

研究目的:

単一系列データを効率的にマイニング

抽出される大量の頻出系列を圧縮

飽和性

- 同じ出現頻度を持つ他の頻出系列の部分系列とならない系列のみを抽出
- すべての頻出系列とその頻度を復元できる

極大性

- 他の頻出系列の部分系列とならない系列のみを抽出
- 頻出系列の頻度情報は復元できない

実験結果(抽出系列数)

DB①

- Webアクセスログデータ
- 系列長:5000
- アイテム系列

最小サポート	頻出系列	頻出右飽和系列	頻出右極大系列
0.04%	75,242	10,731	5,589
0.06%	11,985	5,550	3,347
0.08%	4,563	3,208	2,072
0.1%	2,472	2078	1,375

DB②

- 毎日新聞記事データ1年分
- 系列長:365
- アイテム集合系列

最小サポート	頻出系列	頻出右飽和系列	頻出右極大系列
5%	162,131	162,002	138,161
6%	77,015	76,969	66,088
7%	32,928	32,917	28,459
8%	15,754	15,751	13,648

データベースの特徴により，あまり圧縮できない場合もある

参考:Wang. J, and Han. J, BIDE:Efficient Mining of Frequent Closed Sequences. *Proc. Int. Conf. on Data Engineering (ICDE'04)*,pp.79-90 (2004)