

Übungen zur Vorlesung
Wissensentdeckung in Datenbanken
Sommersemester 2008

Blatt 5

Wiederholung Auch im Rahmen dieses Blattes sollten Sie einige Inhalte aus der Vorlesung wiederholen können. Dabei sollten Sie folgende Fragestellungen beantworten können:

- (a) Was ist ein Verband? Geben Sie einen Beispiel-Verband anhand von Mengen mit der Teilmengenrelation als partieller Ordnung an!
- (b) Der WinEPI-Algorithmus basiert auf diskreten Ereignistypen. Beschreiben Sie das Vorgehen von Gautam Das et. al. um Zeitreihen zu diskretisieren.
- (c) Wie werden zeitliche Phänomene im Ansatz von Frank Höppner dargestellt bzw. gefunden?
- (d) Geben Sie die zentrale Idee des FP-Growth-Algorithmus wieder und beschreiben Sie den Algorithmus kurz.
- (e) Was bewirkt die Sortierung der Itemsets der Transaktionen nach deren Häufigkeit?

Hinweis: Die in der Vorlesung vorgestellten Arbeiten finden Sie unter den Titeln “*Rule discovery from times series*” (Gautam Das et. al., 1998) sowie “*Learning Temporal Rules from State Sequences*” (Frank Höppner, 2001) im Web oder direkt auf der Web-Seite mit den Übungsblättern:

<http://www-ai.cs.uni-dortmund.de/LEHRE/VORLESUNGEN/KDD/SS08/index.html>

Der FP-Growth-Algorithmus ist bekanntermaßen nicht leicht zu verstehen. Bei Verständnisproblemen sei auf das Originalpapier “*Mining Frequent Patterns without Candidate Generation*” (Han et. al., 1999) verwiesen, welches ebenfalls auf der Web-Seite der Übungsblätter zu finden ist.

Aufgabe 5.1

Die folgende Aufgabe behandelt den in der Vorlesung vorgestellten Algorithmus FP-Growth. Als Grundlage dient erneut die Datenbank der Kinogänger (jede Zeile stellt einen Zuschauer dar). Dabei sei ein minimaler Support von $\frac{2}{5}$ gegeben, der für alle nachfolgenden Aufgaben gelten soll.

Bang Boom Bang	Wayne's World	Sissi	Bridget Jones	Otto	E.T.	Indiana Jones	Simpsons	Star Wars
1	0	1	1	1	1	0	1	1
0	1	0	1	0	1	0	0	1
1	0	1	0	0	1	1	0	0
0	1	0	0	1	1	1	0	0
1	0	1	1	0	0	0	1	0
1	0	1	0	1	0	1	0	1
1	0	1	0	1	0	1	0	0
0	0	1	1	0	0	0	0	1
1	0	1	0	0	0	0	0	0
1	0	1	0	0	0	1	0	1

- Geben Sie die Transaktionstabelle mit nach Häufigkeit sortierten Items (innerhalb der Transaktionen) an!
- Bestimmen Sie die Header-Tabelle sowie den *FP-Tree* aus der angegebenen Transaktionstabelle.
- Bestimmen Sie alle *conditional pattern bases* zum *FP-Tree*.
- Bestimmen Sie nun zu den *conditional pattern bases* die *conditional FP-Trees*.
- Bestimmen Sie anhand der *conditional FP-Trees* rekursiv die *frequent patterns*. Zeigen Sie die Erfassung der *frequent patterns* jeweils an der Entwicklung der *conditional pattern bases* sowie den *conditional FP-Trees*.