

Prof. Dr. Kristian Kersting
Prof. Dr. Claus Weihs
Alejandro Molina
Daniel Horn
Leonard Judt

Dortmund, 22.04.14
Abgabe: 30.04.2014 im EWS

Übungen zur Vorlesung
Wissensentdeckung in Datenbanken
Sommersemester 2014
Blatt 2

Wiederholung Im Rahmen dieses Blattes sollten Sie einige Inhalte aus der Vorlesung wiederholen können. Dabei sollten Sie folgende Fragen beantworten können:

1. Was ist ein Verband? Geben Sie einen Beispiel-Verband anhand von Mengen mit der Teilmengenrelation als partieller Ordnung an!
2. Was ist die *Monotonie*-Eigenschaft im Bezug auf häufige Mengen und den *Apriori*-Algorithmus?
3. Geben Sie die zentrale Idee des FP-Growth-Algorithmus wieder und beschreiben Sie den Algorithmus kurz.
4. Was bewirkt die Sortierung der Itemsets der Transaktionen nach deren Häufigkeit?

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

Aufgabe 2.1 (10 Punkte)

In der Vorlesung wurden mit Hilfe des Apriori-Algorithmus die häufigen Mengen in einer Transaktionsdatenbank gefunden. Gegeben sei die nachfolgende Aufstellung von Filmen, die von Zuschauern $z_1, \dots, z_{10}$ besucht worden sind.

1. (2 Punkte) Formen Sie die Tabelle in eine Transaktionsdatenbank um!
2. (8 Punkte) Bestimmen Sie mit dem Apriori-Algorithmus die häufigen Mengen mit minimalem Support von $\frac{2}{5}$ und $\frac{3}{5}$. Geben Sie für jeden Schritt die Kandidatenmenge sowie die Menge der *large itemsets* (d.h. diejenigen Mengen, die den minimalen Support erfüllen) an.

Titel	Jahr	z_1	z_2	z_3	z_4	z_5	z_6	z_7	z_8	z_9	z_{10}
Star Wars	1977	1	1	0	0	1	0	1	0	1	1
E.T. der Ausserirdische	1982	1	1	0	1	1	0	1	0	1	1
Indiana Jones	1989	1	1	1	0	0	0	1	0	1	1
Otto - der Ausserfriesische	1989	0	0	0	0	0	0	1	0	1	1
Wayne's World	1992	1	1	0	1	0	1	0	1	0	1
Bang Boom Bang	1999	1	1	0	1	1	0	0	0	1	1
Bridget Jones	2001	1	0	0	1	0	0	0	1	0	0
Simpsons (Film)	2007	0	0	0	1	1	0	0	0	0	1

Aufgabe 2.2 (10 Punkte)

Diese Aufgabe behandelt den in der Vorlesung vorgestellten Algorithmus FP-Growth. Als Grundlage dient wieder die Datenbank aus Aufgabe 1. Es sei ein minimaler Support von $\frac{2}{5}$ gegeben, für den nun die häufigen Mengen in der Datenbank gefunden werden sollen.

- (2 Punkte) Geben Sie die Transaktionstabelle mit nach Häufigkeit sortierten Items (innerhalb der Transaktionen) an!
- (2 Punkte) Bestimmen Sie die Header-Tabelle sowie den *FP-Tree* aus der angegebenen Transaktionstabelle.
- (2 Punkte) Bestimmen Sie alle *conditional pattern bases* zum *FP-Tree*.
- (2 Punkte) Bestimmen Sie nun zu den *conditional pattern bases* die *conditional FP-Trees*.
- (2 Punkte) 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*.