Prof. Dr. Katharina Morik, Prof. Dr. Claus Weihs
Übungen: Christian Bockermann, Julia Schiffner

Dortmund, den 11. April 2011

Übungen zur Vorlesung

Wissensentdeckung in Datenbanken

Sommersemester 2011
Blatt 1

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 bekanntermaßen 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 1

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 $z_1, \ldots, z_{10}$ besucht worden sind.

Titel	Jahr	z_1	z_2	z_3	z_4	z_5	z_6	z_7	z_8	z_9	z_{10}
Sissi	1955	1	0	1	1	0	0	0	1	0	1
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

1. Formen Sie die Tabelle in eine Transaktionsdatenbank um!
2. 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.

Aufgabe 2

Die in den Übungen vorgestellte Software *RapidMiner* enthält sowohl eine Implementierung des *Apriori*-Algorithmus als auch einen Operator für das *FP-Growth* Verfahren. Die obige Datenbank-Tabelle finden sie als CSV-Datei unter:

`http://kirmes.cs.uni-dortmund.de/wid2011/data/kino.csv`

1. Laden Sie die CSV-Datei herunter und importieren Sie die Daten in ihr *RapidMiner* Repository!
2. Laden Sie die Daten in ein *RapidMiner* Experiment und starten Sie das Experiment. Betrachten Sie die Daten in der Ergebnisansicht von *RapidMiner* und überlegen Sie sich, welche Attribute für den *Apriori*-Algorithmus benötigt werden.
3. Erstellen Sie ein Experiment, dass die Daten liest und den Apriori-Algorithmus auf die Daten anwendet.

Hinweis: *Beachten Sie, dass der* Apriori-*Operator nur binäre Attribute verwenden kann! Sie benötigen dafür einen Operator, der Attribute in das gewünschte Ziel-Format konvertieren kann (vgl. Übungsstunde).*

Aufgabe 3

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.

1. Geben Sie die Transaktionstabelle mit nach Häufigkeit sortierten Items (innerhalb der Transaktionen) an!
2. Bestimmen Sie die Header-Tabelle sowie den *FP-Tree* aus der angegebenen Transaktionstabelle.
3. Bestimmen Sie alle *conditional pattern bases* zum *FP-Tree*.
4. Bestimmen Sie nun zu den *conditional pattern bases* die *conditional FP-Trees*.
5. 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*.