Prof. Dr. Katharina Morik,
Prof. Dr. Claus Weihs
Dipl.-Inform. Klaus Friedrichs,
Dipl.-Stat. Julia Schiffner,
Dr. Issam Ben Khediri

Dortmund, 17.04.12
Abgabe: bis Mi, 25.04., 12.00 Uhr an
friedrichs@statistik.tu-dortmund.de

Übungen zur Vorlesung

Wissensentdeckung in Datenbanken

Sommersemester 2012

Blatt 2

Aufgabe 2.1 (2 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 $z_1, \ldots, z_{10}$ 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}
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
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

Bestimmen Sie mit dem *Apriori*-Algorithmus die häufigen Mengen mit minimalem Support von $s_{\min} = 2/5$. Geben Sie für jeden Schritt k die Kandidatenmenge C_k sowie die Menge L_k der häufigen Mengen an.

Aufgabe 2.2 (3 Punkte)
Diese Aufgabe behandelt den in der Vorlesung vorgestellten Algorithmus *FP-Growth*. Als Grundlage dient wieder die Datenbank aus Aufgabe 2.1. Es sei wieder ein minimaler Support von $s_{\min} = 2/5$ gegeben, für den nun die häufigen Mengen in der Datenbank gefunden werden sollen.

1. Formen Sie die Tabelle in eine Transaktionsdatenbank um und ergänzen Sie sie um eine Spalte mit den *ordered frequent items* für $s_{\min} = 2/5$.
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*.

Aufgabe 2.3 (5 Punkte)
Die in der Vorlesung vorgestellte Software *RapidMiner* enthält sowohl eine Implementierung des *Apriori*-Algorithmus (hierfür bitte die Weka Extension installieren) als auch einen Operator für das *FP-Growth*-Algorithmus.

Auf der Homepage zur Vorlesung in der Datei `groceries.csv` liegen Transaktionsdaten für einen Supermarkt. Die Spalte *customer* enthält die Kunden ID, *item* bezeichnet die Produktgruppe des gekauften Produkts und *amount* die gekaufte Menge.

customer	item	amount
1	citrus fruit	2
1	tropical fruit	1
1	whole milk	2
	⋮	
2	tropical fruit	5
2	root vegetables	3
	⋮	
3	brown bread	4
3	soda	3
	⋮	

1. Laden Sie die .csv-Datei herunter und importieren Sie die Daten in ihr *RapidMiner* Repository! Achten Sie in Step 3 darauf, die erste Zeile als Spaltennamen zu annotieren, und deklarieren Sie in Step 4 die erste Spalte als id. Lesen Sie die Daten mithilfe des *Retrieve*-Operators ein. Starten Sie das Experiment und betrachten Sie die Daten in der Ergebnisansicht von *RapidMiner*, um zu kontrollieren, dass das Einlesen korrekt funktioniert hat.

2. Um die *Apriori* bzw. *FP-Growth*-Operatoren anwenden zu können, müssen Sie zunächst eine Binärdatensatz erstellen. Dieser enthält für jedes item ein Attribut, das angibt, ob es von einem bestimmten Kunden gekauft wurde. Welche Menge genau gekauft wurde, ist hierbei egal. Hier ein Beispiel zu obiger Tabelle:

customer	brown bread	citrus fruit	tropical fruit	whole milk	root vegetables	soda	...
1	false	true	true	true	false	false	...
2	false	false	true	false	true	false	...
3	true	false	false	false	false	true	...
			⋮				

   Nützliche Operatoren zur Transfomierung des Datensatzes sind *Pivot*, *Replace Missing Values* und *Numerical to Binominal*.

3. Wenden Sie den *Apriori*-Algorithmus an. Ändern Sie die voreingestellten Parameter und schauen Sie, wie sich die Einstellungen auf die gefundenen häufigen Mengen und die Assoziationsregeln auswirken.

4. Finden Sie häufige Mengen mithilfe des *FP-Growth*-Verfahrens und generieren Sie daraus Assoziationsregeln. Spielen Sie auch hier mit verschiedenen Parametereinstellungen.

5. Bauen Sie einen weiteren Operator, z.B. *Item Sets to Data* oder *Generate Item Set Indicators*, sinnvoll in Ihr Experiment ein, und beschreiben Sie kurz, was er macht.