Prof. Dr. Katharina Morik, JProf. Dr. Uwe Ligges
Christian Bockermann, Gero Szepannek

Dortmund, 17. April

Abgabe: bis Dienstag, 29.04., 23:59h
an christian.bockermann@cs.uni-dortmund.de

Übungen zur Vorlesung
Wissensentdeckung in Datenbanken
Sommersemester 2008

Blatt 3

Aufgabe 3.1
In der Vorlesung wurden mit Hilfe des Apriori-Algorithmus die häufigen Mengen einer Transaktionsdatenbank gefunden. Auf Basis dieser häufigen Mengen sind dann Assoziationsregeln generiert worden.

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

(a) Bestimmen Sie mit Hilfe des Apriori-Algorithmus die häufigen Mengen mit minimalem Support von 0.4 und 0.6. Geben Sie dabei für jeden Schritt die Kandidatenmenge sowie die Menge der *large itemsets* (d.h. diejenigen Mengen, die den minimalen Support erfüllen) an.

(b) Geben Sie alle Regeln mit minimalem Support von 0.4 und einer Konfidenz von mindestens 0.8 an.

Aufgabe 3.2
Sei $r = A \rightarrow B$ eine Regel und $\text{conf}(r) = \text{conf}(A \rightarrow B)$ die Konfidenz von r. Der Support von r werde mit $\text{supp}(r) = \text{supp}(A \rightarrow B)$ bezeichnet. Die Anzahl der Transaktionen, in denen eine Menge M vorkommt, werde mit $h(M)$ bezeichnet.

Zeigen Sie für folgende Aussagen, ob sie gelten oder nicht. Geben Sie dabei immer einen Beweis oder ein Gegenbeispiel an.

(a) $((\text{conf}(A \rightarrow B) = \alpha) \wedge \text{conf}(B \rightarrow C) = \beta) \Rightarrow \text{conf}(A \rightarrow C) = \alpha\beta$

(b) $\text{conf}(A \rightarrow B) = \text{conf}(B \rightarrow A) \Rightarrow h(A) = h(B)$

(c) $\text{supp}(X \rightarrow Y) \geq \text{supp}(X \rightarrow \emptyset) \cdot \text{supp}(Y \rightarrow \emptyset)$

Aufgabe 3.3
In einer kleinen Datenbank seien folgende Relationen *FilmStudio* und *FilmStar* gegeben:

Title	Jahr	Studio
Sissi	1955	Erma-Film
Star Wars	1977	Lucasfilm
E.T. der Außerirdische	1982	Universal
Indiana Jones	1989	Lukas Film
Otto - der Außerfriesische	1989	Rialto Film
Wayne's World	1992	Paramount
Bang Boom Bang	1999	Becker & Häberle
Bridget Jones	2001	Little Bird
Simpsons (Film)	2007	Fox
⋮	⋮	⋮

Title	Jahr	Star
Sissi	1955	Romy Schneider
Star Wars	1977	C.Fisher
Star Wars	1977	H. Ford
Star Wars	1977	M.Hamill
E.T. der Außerirdische	1982	H. Thomas
E.T. der Außerirdische	1982	D. Barrymore
E.T. der Außerirdische	1982	E. Eleniak
Indiana Jones	1989	H. Ford
Indiana Jones	1989	S. Connery
⋮	⋮	⋮

Dazu existiert noch eine weitere Relation *Bewertung*, die neben "titel" und "jahr" die Attribute "bewertung" und "user" enthält. Jedes Tupel in dieser Relation stellt also die Bewertung eines Filmes durch einen Benutzer dar.

Geben Sie für die folgenden Fragestellungen ein *kommentiertes*[1] SQL-Statement an.

(a) Mit welchen Produktionsfirmen (Studio) haben die Stars zusammengearbeitet? Ihre Anfrage sollte eine Menge von Tupeln zurückliefern, die einem Studio einen Star zuordnet.

(b) Welche durchschnittlichen Bewertungen haben die Filme bekommen?
(Hinweis: SQL erlaubt arithmetische Ausdrücke, z.B. `SELECT 2 + 3`)

(c) Geben Sie eine Anfrage an, die für jede Produktionsfirma den Durchschnitt der Bewertungen der Filme berechnet, die das Studio produziert hat.

Um ihre Lösungen auszuprobieren, finden Sie eine kleine Datenbank mit diesen Relationen und einigen Einträgen als SQLite-Datenbank Datei auf den Seiten zur Vorlesung unter:

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

[1] Kommentiert soll in diesem Fall bedeuten, dass Sie eine **kurze** textuelle Beschreibung zu ihrer SQL-Anfrage geben, die erklärt wie Sie die Anfrage zusammengesetzt haben.