Dr. Andreas Hotho, Robert Jäschke
Fachgebiet Wissensverarbeitung

11.12.2008

7. Übung „Knowledge Discovery“

Wintersemester 2008/2009

1 Häufige Itemmengen

Beweisen Sie die Korrektheit der folgenden zur Regelgenerierung verwendeten Abhängigkeit für häufige Itemmengen X und deren Teilmengen $\hat{X} \subset X$. Verwenden Sie dazu die textuelle Beschreibung der Konfidenz im Kapitel 4.2 auf Folie 6.

$$\text{confidence}\left((X \setminus \hat{X}) \rightarrow \hat{X}\right) = \frac{\text{support}(X)}{\text{support}(X \setminus \hat{X})}$$

$$\text{confidence}\left((X \setminus \hat{X}) \rightarrow \hat{X}\right) = \frac{|\{t \in D \mid (X \setminus \hat{X}) \cup \hat{X} \subseteq t\}|}{|\{t \in D \mid (X \setminus \hat{X}) \subseteq t\}|} \tag{1}$$

$$= \frac{|\{t \in D \mid X \subseteq t\}|}{|\{t \in D \mid (X \setminus \hat{X}) \subseteq t\}|} \tag{2}$$

$$= \frac{|\{t \in D \mid X \subseteq t\}|}{|D|} \frac{|D|}{|\{t \in D \mid (X \setminus \hat{X}) \subseteq t\}|} \tag{3}$$

$$= \frac{\text{support}(X)}{\text{support}(X \setminus \hat{X})} \tag{4}$$

2 Assoziationsregeln

Gegeben sei die folgende Menge D von Warenkorbdaten (Transaktionen):

Transaktion	Items
t_1	Windeln, Bier, Chips
t_2	Chips, TV-Zeitschrift
t_3	TV-Zeitschrift, Bier, Chips
t_4	Bier, Windeln, Zahnpasta
t_5	Zahnpasta, Chips
t_6	TV-Zeitschrift, Chips, Bier
t_7	Bier, Windeln
t_8	TV-Zeitschrift, Chips

1. Bestimmen Sie zu den gegebenen Transaktionen die häufigen Itemmengen, die einen Mindestsupport von 25% aufweisen. Gehen Sie bei der Bestimmung der häufigen Itemmengen nach dem aus der Vorlesung bekannten Algorithmus vor.

C_k**:** die zu zählenden Kandidaten-Itemsets der Länge k

L_k**:** Menge aller häufig vorkommenden Itemsets der Länge k

Apriori$(I, D, minsup)$
$L_1 := \{$frequent 1-Itemsets aus $I\}$;
$k := 2$;
while $L_{k-1} \neq \emptyset$ **do**
$C_k := \text{AprioriKandidatenGenerierung}(L_{k-1})$;
for each Transaktion $T \in D$ **do**
$CT := \text{Subset}(C_k, T)$; // alle Kandidaten aus C_k,
// die in der Transaktion T enthalten sind
for each Kandidat $c \in CT$ **do** c.count $++$;
$L_k := \{c \in C_k \mid (c.\text{count}/|D|) \geq minsup\}$;
$k++$;
return $\bigcup_k L_k$;

1. Durchlauf:

Item	Transaktionen, die das Item enthalten	Support des Items
Chips	$t_1, t_2, t_3, t_5, t_6, t_8$	**6/8 = 75%**
TV-Zeitschrift	t_2, t_3, t_6, t_8	**4/8 = 50%**
Bier	t_1, t_3, t_4, t_6, t_7	**5/8 = 62.5%**
Windeln	t_1, t_4, t_7	**3/8 = 37.5%**
Zahnpasta	t_4, t_5	**2/8 = 25%**

a) $minsup = 25\%, k := 1; L_1 := \{\}$;
$C_1 := \{\{\text{Chips}\}, \{\text{TV-Zeitschrift}\}, \{\text{Bier}\}, \{\text{Windeln}\}, \{\text{Zahnpasta}\}\}$;

b) Support für alle $L \in L_k$ bestimmen (siehe 1. Tabelle)

c) $L_1 := \{\{\text{Chips}\}, \{\text{TV-Zeitschrift}\}, \{\text{Bier}\}, \{\text{Windeln}\}, \{\text{Zahnpasta}\}\}$;
$L := \{\{\text{Chips}\}, \{\text{TV-Zeitschrift}\}, \{\text{Bier}\}, \{\text{Windeln}\}, \{\text{Zahnpasta}\}\}$;

d) $L_1 \neq \emptyset$;

e) C_2 :={{Chips, TV-Zeitschrift}, {Chips, Bier}, {Chips, Windeln}, {Chips, Zahnpasta}, {TV-Zeitschrift, Bier}, {TV-Zeitschrift, Windeln}, {TV-Zeitschrift, Zahnpasta}, {Bier, Windeln}, {Bier, Zahnpasta}, {Windeln, Zahnpasta}};
$k := k + 1 = 2$;

f) Gehe nach b.

2. Durchlauf:

Item	Transaktionen, die das Item enthalten	Support des Items
Chips, TV-Zeitschrift	t_2, t_3, t_6, t_8	**4/8 = 50%**
Chips, Bier	t_1, t_3, t_6	**3/8 = 37.5%**
Chips, Windeln	t_1	1/8 = 12.5%
Chips, Zahnpasta	t_5	1/8 = 12.5%
TV-Zeitschrift, Bier	t_3, t_6	**2/8 = 25%**
TV-Zeitschrift, Windeln	-	0%
TV-Zeitschrift, Zahnpasta	-	0%
Bier, Windeln	t_1, t_4, t_7	**3/8 = 37.5%**
Bier, Zahnpasta	t_4	1/8 = 12.5%
Windeln, Zahnpasta	t_4	1/8 = 12.5%

b) siehe 2. Tabelle

c) L_2 :={{Chips,TV-Zeitschrift}, {Chips, Bier}, {TV-Zeitschrift, Bier}, {Bier,Windeln}}
L :={{Chips}, {TV-Zeitschrift}, {Bier}, {Windeln}, {Zahnpasta}, {Chips, TV-Zeitschrift}, {Chips, Bier}, {TV-Zeitschrift, Bier}, {Bier, Windeln}}

d) $L_2 \neq \emptyset$

e) C_3 :={{Chips, TV-Zeitschrift, Bier}}
$k := k + 1 = 3$

f) Gehe nach b.

3. Durchlauf:

Item	Transaktionen, die das Item enthalten	Support des Items
Chips, TV-Zeitschrift, Bier	t_3, t_6	**2/8 = 25%**

Jetzt: Abbruch, da $L_4 = \emptyset$.
Ergebnis: Folgende Itemmengen sind häufig mit einem Mindestsupport von 25%:
I = {{Chips}, {TV-Zeitschrift}, {Bier}, {Windeln}, {Zahnpasta}, {Chips, TV-Zeitschrift}, {Chips, Bier}, {TV-Zeitschrift, Bier}, {Bier, Windeln}, {Chips, TV-Zeitschrift, Bier}}

2. Bestimmen Sie nun aus den berechneten Itemmengen alle Assoziationsregeln mit einer Mindestkonfidenz von 66%.

Regel	Konfidenz
Chips $\Longrightarrow$ TV-Zeitschrift	**66.67%**
TV-Zeitschrift $\Longrightarrow$ Chips	**100%**
Chips $\Longrightarrow$ Bier	50%
Bier $\Longrightarrow$ Chips	60%
TV-Zeitschrift $\Longrightarrow$ Bier	50%
Bier $\Longrightarrow$ TV-Zeitschrift	40%
Bier $\Longrightarrow$ Windeln	60%
Windeln $\Longrightarrow$ Bier	**100%**
Chips $\Longrightarrow$ TV-Zeitschrift, Bier	33.34%
TV-Zeitschrift $\Longrightarrow$ Chips, Bier	50%
Bier $\Longrightarrow$ TV-Zeitschrift, Chips	40%
TV-Zeitschrift, Bier $\Longrightarrow$ Chips	**100%**
Chips, Bier $\Longrightarrow$ TV-Zeitschrift	**66.67%**
TV-Zeitschrift, Chips $\Longrightarrow$ Bier	50%

3. Wie können die erzielten Ergebnisse interpretiert werden? Wie sähe die Interpretation aus, wenn statt dessen 20000 Transaktionen als Grundlage des Assoziationsregel-Mining gedient hätten?

   - Für eine Interpretation ist die Datenbasis zu klein.
   - Regeln mit 50% Konfidenz sind in diesem Falle nicht aussagekräftig.
   - 100% Regeln sind hier meistens trivial.

   Anwendung des Assoziationsregel-Verfahrens auf einer großen Anzahl von Transaktionen:

   - Triviale und offensichtliche Regeln werden entdeckt (100%-Regeln); 50-66%-Regeln können neue Informationen enthalten.
   - Tausende von Regeln werden generiert.
   - Problem: wie relevante Regeln ausfiltern bzw. „entdecken“?

4. Ein weiteres Maß für Regeln ist der sogenannte *Lift* einer Regel, welcher definiert ist durch

$$\text{lift}(X \rightarrow Y) = \frac{\text{confidence}(X \rightarrow Y)}{\text{support}(Y)}$$

   Was besagt der Lift einer Regel?

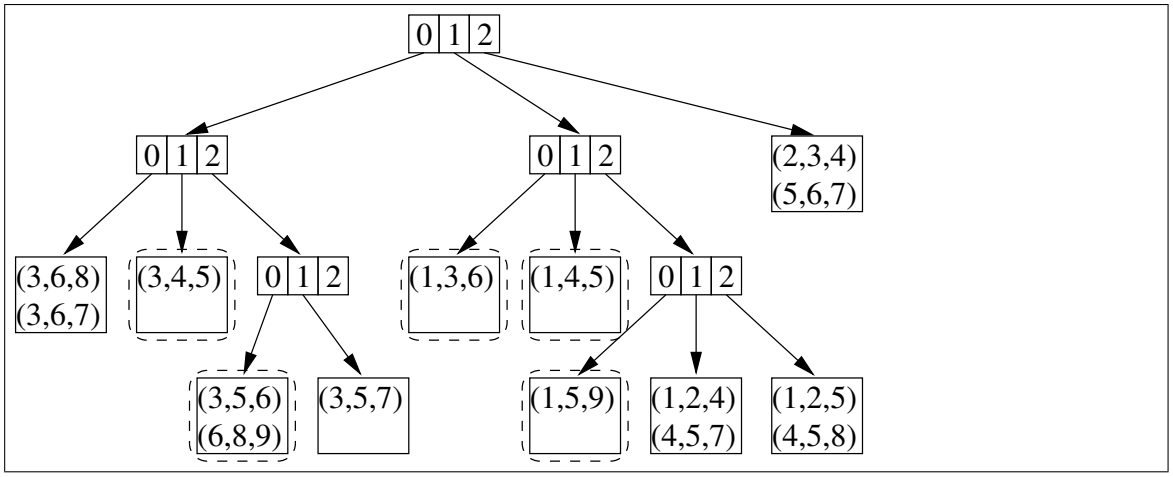
   Je größer der Lift von $X \rightarrow Y$, umso größer ist der Einfluß der Itemmenge X auf die Wahrscheinlichkeit des Auftretens von Itemmenge Y.

   Beispiel:
   Hammer $\rightarrow$ Nägel (30 %; 8 %): 3.75
   Hammer, Nägel $\rightarrow$ Bauholz (33 %; 2 %): 16.5

3 Hashbaum

1. Konstruieren Sie einen Hashbaum für die folgenden Itemsets für $k = 3$ und maximal 2 Transaktionen pro Knoten.

   (1,4,5), (2,3,4), (3,6,8), (1,2,5), (4,5,8), (3,4,5), (1,2,4), (1,3,6), (3,5,6), (4,5,7), (3,5,7), (1,5,9), (6,8,9), (3,6,7), (5,6,7)



2. Welche Blattknoten müssen besucht werden, wenn die 3-Item Teilmengen von (1,3,7,8,9) gefunden werden sollen?

   Die entsprechenden Blattknoten sind im obigen Graphen mit einer gestrichelten Linie umrahmt.