Universität Kassel
Fachbereich Mathematik/Informatik
Fachgebiet Wissensverarbeitung
Hertie-Stiftungsprofessur
Wilhelmshöher Allee 73
34121 Kassel

Email: hotho@cs.uni-kassel.de
Tel.: ++49 561 804-6252

Dr. Andreas Hotho
Prof. Gerd Stumme

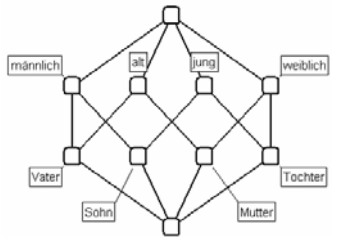
14.07.04

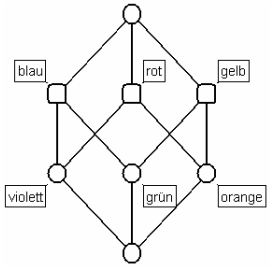
Knowledge Discovery

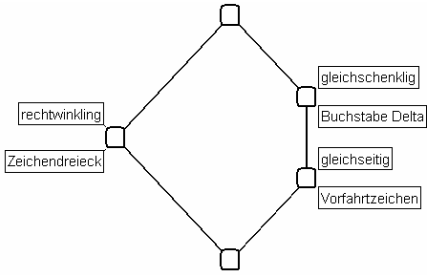
Lösungsblatt 6

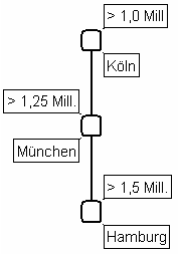
Sommersemester 2004

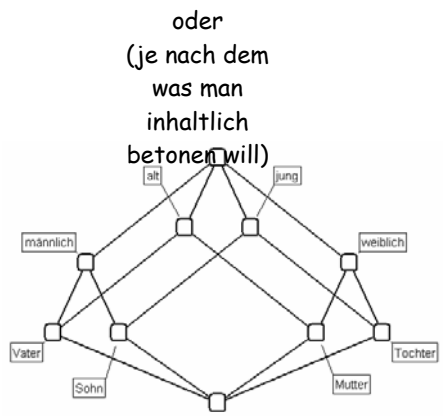
Aufgabe 3

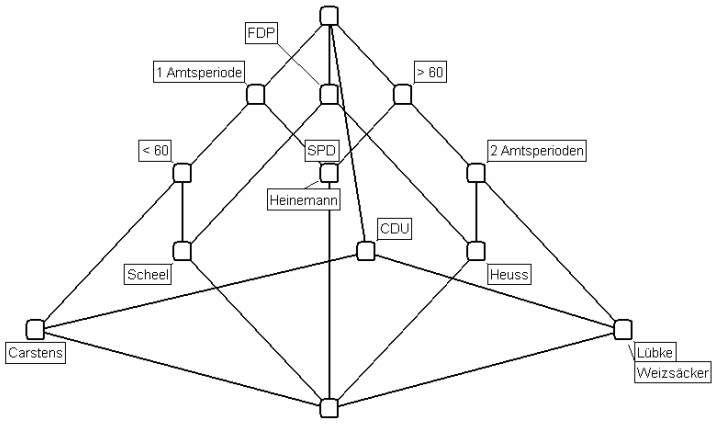












Aufgabe 4 (FCA)

Seien (G,M,I) ein formaler Kontext und $A_1, A_2 \subseteq G$. Zeigen Sie, dass folgendes gilt:

$A_1 \subseteq A_2 \Rightarrow A_2' \subseteq A_1'$

Beweis:

Sei $A_1 \subseteq A_2$ und $m \in A_2'$.

Zu zeigen: $m \in A_1'$.

Wegen $m \in A_2'$ gilt für alle $g \in A_2$, dass $(g,m) \in I$. Da $A_1 \subseteq A_2$, gilt auch für alle $g \in A_1$, dass $(g,m) \in I$. Also $m \in \{n \in M \mid \forall g \in A_1 : (g,m) \in I\} = A_1'$.

AIFB

Apriori-Algorithmus zur Berechnung häufiger Itemmengen:

1. Initialisierung:

   s_{min} = Wert für minimalen Support;
   n := 1;
   I := $\emptyset$;
   H_n := {{i} | i ist ein Item};

2. gehe über die Datenbasis D und bestimme für alle $H \in H_n$ den Support;

3. $I_n := \{H \in H_n \mid support(H) \geq s_{min}\}$;
   $I := I \cup I_n$;

4. Falls $I_n = \emptyset$, gebe I als Ergebnis aus;

5. $H_{n+1} := \{\{i_1,i_2,...,i_{n+1}\} \mid \forall j: 1 \leq j \leq n+1:(\{i_1,i_2,...,i_{n+1}\} - \{i_j\}) \in I_n\}$;
   n := n+1;

6. gehe nach 2.

AIFB

Aufgabe 1

a)

$$\text{confidence}((X - X') \rightarrow X') = \frac{\left|\{t \in D \mid [(X - X') \cup X'] \subseteq t\}\right|}{\left|\{t \in D \mid (X - X') \subseteq t\}\right|}$$

$$= \frac{\left|\{t \in D \mid X \subseteq t\}\right|}{\left|\{t \in D \mid (X - X') \subseteq t\}\right|}$$

$$= \frac{\left|\{t \in D \mid X \subseteq t\}\right|}{\left|D\right|} \cdot \frac{\left|D\right|}{\left|\{t \in D \mid (X - X') \subseteq t\}\right|}$$

$$= \frac{\text{support}(X)}{\text{support}(X - X')}$$

Musterlösung Übung 6 -- Aufgabe 2

AIFB

a) Bestimmen sie die häufigen Itemmengen, die einen Mindest-support von 25% aufweisen.

Tabelle I:

Item	Transaktionen, die das Item enthalten	Support des Items
Chips	t1, t2, t3, t5, t6, t8	6/8=75%
TV-Zeitschrift	t2, t3, t6, t8	4/8=50%
Bier	t1, t3, t4, t6, t7	5/8=62.5%
Windeln	t1, t4, t7	3/8=37.5%
Zahnpasta	t4, t5	2/8=25%

AIFB

Anwendung des Apriori-Algorithmus

1. s_min = 25%; n:=1; I:={};
   H_1 := {{Chips}, {TV-Zeitschrift}, {Bier}, {Windeln}, {Zahnpasta}};

2. support für alle H $\in$ H_n bestimmen (siehe Tabelle I);

3. I_1 := {{Chips}, {TV-Zeitschrift}, {Bier}, {Windeln}, {Zahnpasta}};
   I:= {{Chips}, {TV-Zeitschrift}, {Bier}, {Windeln}, {Zahnpasta}};

4. I_1 $\neq \emptyset$;

5. H_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}};
   n := n+1 = 2;

6. Gehe nach 2.

Musterlösung Übung 6 -- Aufgabe 2

2. Durchlauf liefert für Schritt 2 (Tabelle II):

Itemmenge	Transaktionen, die die Itemmenge enthalten	Support der Itemmenge
Chips, TV-Zeitschrift	t2, t3, t6, t8	**4/8=50%**
Chips, Bier	t1, t3, t6	**3/8=37.5%**
Chips, Windeln	t1	1/8=12.5%
Chips, Zahnpasta	t5	1/8=12.5%
TV-Zeitschrift, Bier	t3, t6	**2/8=25%**
TV-Zeitschrift, Windeln	-	0%
TV-Zeitschrift, Zahnpasta	-	0%
Bier, Windeln	t4, t7	**2/8=25%**
Bier, Zahnpasta	t4	1/8=12.5%
Windeln, Zahnpasta	t4	1/8=12.5%

Musterlösung Übung 6 -- Aufgabe 2

AIFB

2. Durchlauf:

2. vgl. Tabelle II

3. I_2 := {{Chips, TV-Zeitschrift}, {Chips, Bier}, { TV-Zeitschrift, Bier}
{Bier, Windeln}};
I := {{Chips}, {TV-Zeitschrift}, {Bier}, {Windeln}, {Zahnpasta},
{Chips, TV-Zeitschrift}, {Chips, Bier}, {TV-Zeitschrift, Bier}
{Bier, Windeln}};

4. I_2 $\neq \emptyset$;

5. H_3 := {{Chips, TV-Zeitschrift, Bier},{Bier,Windeln,TV-Zeitschrift},
{Bier,Windeln,Chips}}
n := n+1 = 3;

6. gehe nach 2.

Musterlösung Übung 6-- Aufgabe 2

AIFB

3. Durchlauf:

2. vgl. Tabelle III

3. I_3 := {{Chips, TV-Zeitschrift, Bier};
   I := {{Chips}, {TV-Zeitschrift}, {Bier}, {Windeln}, {Zahnpasta},
   {Chips, TV-Zeitschrift}, {Chips, Bier}, { TV-Zeitschrift, Bier}
   {Bier, Windeln}, {Chips, TV-Zeitschrift, Bier}};

4. I_3 $\neq \emptyset$;

5. H_4 = $\emptyset$
   n := n+1 = 4;

6. gehe nach 2.

Musterlösung Übung 6 -- Aufgabe 2

3. Durchlauf liefert für Schritt 2 (Tabelle III):

Itemmenge	Transaktionen, die die Itemmenge enthalten	Support der Itemmenge
Chips, TV-Zeitschrift, Bier	t3, t6	**2/8=25%**
Bier, Windeln, Chips	t1	**1/8=12.5%**
Bier, Windeln, TV-Zeit.	/	**0%**

Jetzt: Abbruch, da I_4 = ∅:

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}}

AIFB

Aufgabe 2

b) Bestimmen sie alle Assoziationsregeln mit einer Mindestkonfidenz von 66%:

Regel	Konfidenz
Chips => TV-Zeitschrift	**66.67%**
TV-Zeitschrift => Chips	**100%**
Chips => Bier	50%
Bier => Chips	60%
TV-Zeitschrift => Bier	50%
Bier => TV-Zeitschrift	40%
Bier => Windeln	40%
Windeln => Bier	**66.67%**
Chips => TV-Zeitschrift, Bier	33.34%
TV-Zeitschrift => Chips, Bier	50%
Bier => TV-Zeitschrift, Chips	40%
TV-Zeitschrift, Bier => Chips	**100%**
Chips, Bier => TV-Zeitschrift	**66.67%**
TV-Zeitschrift, Chips => Bier	50%

AIFB

Aufgabe 2

c) Lift einer Regel:

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

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

AIFB

Aufgabe 2

d) Interpretation der Ergebnisse:

- In diesem Fall ist Datenbasis für Aussagen zu klein
- Regeln mit 50% Konfidenz sind in diesem nicht aussagekräftig.
- 100% Regeln sind hier im Regelfall trivial

Anwendung des Assoziationsregelverfahrens auf einer großen Anzahl von Transaktionen:

- triviale und offensichtliche Regel werden entdeckt (100% - Regeln); 50-66% Regeln können neue Informationen enthalten
- Tausende von Regeln werden generiert.
- Problem: Wie relevante Regeln ausfiltern bzw. ‘‘entdecken‘‘?