Wissensentdeckung mit Assoziationsregeln

- verständliche Form $\longrightarrow$ Assoziationsregeln

$$r : A_1 \ldots A_n \rightarrow B_1 \ldots B_m$$

Beispiel:

"80 % der Kunden, die Bier und Sekt kaufen, kaufen auch Kartoffelchips"

- nützlich und interessant $\longrightarrow$ Wahrscheinlichkeiten nahe 1
  - Support $support(r) = Prob(A_1 \ldots A_n, B_1 \ldots B_m)$
  - Konfidenz $confidence(r) = Prob(B_1 \ldots B_m | A_1 \ldots A_n)$

Finde Assoziationsregeln mit hohem Support und hoher Konfidenz!

Wissensentdeckung

- Item = Artikel, Ding, Gegenstand;
  $\mathcal{I}$ Menge aller betrachteten Items
- $X \subseteq \mathcal{I}$ Itemmenge; $|X| = k$: k-Itemmenge
- Transaktion $t \subseteq \mathcal{I}$ Itemmenge;
  Datenbasis $\mathcal{D} = \{t_1, t_2, \ldots\}$ Menge von Transaktionen
- Support einer Itemmenge

$$support(X) = \frac{|\{t \in \mathcal{D} \mid X \subseteq t\}|}{|\mathcal{D}|}$$

Wissensentdeckung

Anwendungsbereich: Verkaufsdatenanalyse eines Supermarktes

Items: Artikel des Marktsortimentes

Transaktionen: Kundeneinkäufe

Datenbasis $\mathcal{D}$: alle Verkaufstransaktionen eines bestimmten Zeitraums

Support der Itemmenge $\{Milch\}$: Prozentsatz der Kunden, die bei ihrem Einkauf auch Milch einkauften.

Wissensentdeckung

Eine Assoziationsregel hat die Form

$$X \rightarrow Y$$

wobei X und Y disjunkte Itemmengen sind.

Der Support einer Assoziationsregel ist definiert als der Support von Prämisse + Konklusion:

$$support(X \rightarrow Y) = support(X \cup Y)$$

Die Konfidenz einer Assoziationsregel $X \rightarrow Y$ ist der (relative) Anteil derjenigen der X enthaltenden Transaktionen, die auch Y enthalten:

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

Wissensentdeckung

Finde alle Assoziationsregeln, die in der betrachteten Datenbasis mit einem Support von mindestens *minsupp* und einer Konfidenz von mindestens *minconf* gelten, wobei *minsupp* und *minconf* benutzerdefinierte Werte sind.

- Finde alle Itemmengen, deren Support über der *minsupp*-Schwelle liegt; diese Itemmengen werden häufige Itemmengen (frequent itemsets) genannt.
- Finde in jeder häufigen Itemmenge I alle Assoziationsregeln

$$I' \rightarrow (I - I') \text{ mit } I' \subset I,$$

deren Konfidenz mindestens *minconf* beträgt.

Wissensentdeckung

$|\mathcal{I}| = n \rightarrow 2^n$ Teilmengen sind zu untersuchen

Idee des Apriori-Algorithmus basiert auf:

> Alle Teilmengen einer häufigen Itemmenge sind ebenfalls häufig, und alle Obermengen einer nicht-häufigen Itemmenge sind ebenfalls nicht häufig.

Dies ist klar, denn für zwei Itemmengen I_1, I_2 gilt:

$$I_1 \subseteq I_2 \text{ impliziert } \mathit{support}(I_2) \leq \mathit{support}(I_1)$$

Wissensentdeckung

Aus den häufigen Itemmengen müssen noch die gesuchten Assoziationsregeln mit einer Konfidenz $\geq minconf$ bestimmt werden. Dabei nutzt man folgenden Zusammenhang aus:

> Beträgt für Itemmengen X, Y mit $Y \subset X$ die Konfidenz einer Regel $(X - Y) \rightarrow Y$ mindestens $minconf$, so gilt dies auch für jede Regel der Form $(X - Y') \rightarrow Y'$ mit $Y' \subseteq Y$.

Erfüllt also eine Assoziationsregel das Konfidenzkriterium, so auch alle Regeln, die sich aus denselben Items und mit kürzerer Konklusion bilder lassen.

$\rightarrow$ Bilde zuerst Assoziationsregeln mit möglichst kurzer Konklusion und erweitere die Konklusion schrittweise.

Wissensentdeckung

Einkaufstransaktionen in einem Drogeriemarkt:

Label	*Artikel*	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	*support*
A	Seife	•				•		•		•		0.4
B	Shampoo	•	•	•	•		•		•	•	•	0.8
C	Haarspülung		•	•	•		•		•	•		0.6
D	Duschgel	•			•		•	•		•	•	0.6
E	Zahnpasta	•		•		•		•				0.4
F	Zahnbürste			•		•						0.2
G	Haarfärbemittel		•		•				•			0.3
H	Haargel		•									0.1
J	Deodorant			•	•	•	•	•	•			0.6
K	Parfüm						•		•			0.2
L	Kosmetikartikel		•		•		•		•		•	0.5

Wissensentdeckung

Festlegung des minimalen Supports und der minimalen Konfidenz:

$$minsupp = 0.4, \quad minconf = 0.7$$

In realen Anwendungen wird $minsupp$ in der Regel sehr viel kleiner gewählt (oft $< 1\%$).

Bestimmung der häufigen k-Itemmengen:
k=1:

$$L_1 = \{\{A\}, \{B\}, \{C\}, \{D\}, \{E\}, \{J\}, \{L\}\}$$

k=2:
Berechnung der Kandidatenmenge C_2 für L_2 durch paarweise Kombinationen der Mengen in L_1:

Wissensentdeckung

C_2-Menge	Support	C_2-Menge	Support	C_2-Menge	Support
{A,B}	0.2	{B,D}	0.5	{C,L}	0.4
{A,C}	0.1	{B,E}	0.2	{D,E}	0.2
{A,D}	0.2	{B,J}	0.4	{D,J}	0.3
{A,E}	0.3	{B,L}	0.5	{D,L}	0.3
{A,J}	0.2	{C,D}	0.3	{E,J}	0.3
{A,L}	0.0	{C,E}	0.1	{E,L}	0.0
{B,C}	0.6	{C,J}	0.4	{J,L}	0.3

Bemerkung: Kein Teilmengencheck, da per Konstruktion alle betrachteten 1-Teilmengen auch in L_1 liegen.

Wissensentdeckung

Also ist

$$L_2 = \{\{B, C\}, \{B, D\}, \{B, J\}, \{B, L\}, \{C, J\}, \{C, L\}\}$$

k=3:

Berechnung der Kandidatenmenge C_3 (zum Vergleich mit und ohne Teilmengencheck, d.h. alle 2-Teilmengen müssen in L_2 liegen):

C_3 *vor Teilmengencheck*	C_3 *nach Teilmengencheck*	*Support*
{B,C,D}	{B,C,J}	0.4
{B,C,J}	{B,C,L}	0.4
{B,C,L}		
{B,D,J}		
{B,D,L}		
{B,J,L}		
{C,J,L}		

Wissensentdeckung

Damit ist

$$L_3 = \{\{B, C, J\}, \{B, C, L\}\}$$

Die einzig mögliche weitere Kombination $\{B, C, J, L\}$ ist nicht häufig, da (z. B.) $\{C, J, L\}$ nicht in L_3 enthalten ist; folglich ist

$$C_4 = L_4 = \emptyset$$

Berechnung der Assoziationsregeln aus den häufigen Itemmengen:
Es bezeichne (wie oben) H_m die Menge der m-Item-Konklusionen der jeweils betrachteten häufigen Itemmenge.

Wissensentdeckung

Regeln der Länge 2, d.h. aus

$$L_2 = \{\{B, C\}, \{B, D\}, \{B, J\}, \{B, L\}, \{C, J\}, \{C, L\}\}$$

gebildete Regeln:

Regel	*Konfidenz*	*Regel*	*Konfidenz*
$B \rightarrow C$	0.75	$C \rightarrow B$	1.00
$B \rightarrow D$	0.63	$D \rightarrow B$	0.83
$B \rightarrow J$	0.50	$J \rightarrow B$	0.67
$B \rightarrow L$	0.63	$L \rightarrow B$	1.00
$C \rightarrow J$	0.67	$J \rightarrow C$	0.67
$C \rightarrow L$	0.67	$L \rightarrow C$	0.80

Wissensentdeckung

Aus $\{B, C, L\}$ erhält man die Regeln

$$BC \rightarrow L\ [0.67], \quad BL \rightarrow C\ [0.8], \quad CL \rightarrow B\ [1.00]$$

und durch Erweiterung der Konklusion noch

$$L \rightarrow BC\ [0.8]$$

die ebenfalls ausgegeben wird.

Durch den Apriori-Algorithus werden insgesamt folgende Regeln berechnet:

Wissensentdeckung

Regel		*Support*	*Konfidenz*
Shampoo	→ Haarspülung	0.6	0.75
Haarspülung	→ Shampoo	0.6	1.00
Duschgel	→ Shampoo	0.5	0.83
Kosmetik	→ Shampoo	0.5	1.00
Kosmetik	→ Haarspülung	0.4	0.80
Shampoo, Deodorant	→ Haarspülung	0.4	1.00
Haarspülung, Deodorant	→ Shampoo	0.4	1.00
Shampoo, Kosmetik	→ Haarspülung	0.4	0.80
Haarspülung, Kosmetik	→ Shampoo	0.4	1.00
Kosmetik	→ Shampoo, Haarspülung	0.4	0.80