Dr. Andreas Hotho, Robert Jäschke
Fachgebiet Wissensverarbeitung

18.12.2008

8. Übung „Knowledge Discovery“

Wintersemester 2008/2009

1 Hierarchische Assoziationsregeln

1. Seien A, A_1, A_2, B Items und A eine Verallgemeinerung von A_1 und A_2. Beweisen oder widerlegen Sie (durch Angabe eines Gegenbeispiels) die folgenden Behauptungen:

   a) $\text{support}(A \to B) = \text{support}(A_1 \to B) + \text{support}(A_2 \to B)$

   Diese Implikation gilt nicht. Ein Gegenbeispiel bilden die Transaktionen $D = \{(A_1, A_2, B), (A_1, B), (A_2, B)\}$ für die gilt:

   $\text{support}(A \to B) = 1 \neq 4/3 = 2/3{+}2/3 = \text{support}(A_1 \to B){+}\text{support}(A_2 \to B)$

   b) Falls $\text{support}(A_1 \to B) > minsupport$, dann ist $\text{support}(A \to B) > minsupport$.

   Gilt, denn

   $$\begin{aligned}\text{support}(A \to B) &= \frac{|(A \cup B)|}{|D|} = \frac{|(A_1 \cup A \cup B)|}{|D|} \\ &\geq \frac{|(A_1 \cup B)|}{|D|} = \text{support}(A_1 \to B) > minsupport\end{aligned}$$

   c) Falls $\text{support}(A \to B) > minsupport$, dann ist $\text{support}(A_1 \to B) > minsupport$.

   Diese Richtung gilt nicht, wie man für $minsupport = 0.8$ und das in 1a gegegebene D leicht sieht. Dort ist $\text{support}(A \to B) = 1$ aber $\text{support}(A_1 \to B) = 2/6 < 0.8$.

2. Wie berechnet sich die erwartete Konfidenz einer Regel $X \to Y$?

$$\text{erwarteteKonfidenz}(X \to Y) = \frac{\text{support}(X, Y)}{\text{support}(X)}$$

wären X und Y statistisch unabhängig voneinander, so wäre

$$\text{support}(X, Y) = \text{support}(X)\,\text{support}(Y)$$

und damit

$$\text{erwarteteKonfidenz}(X \to Y) = \frac{\text{support}(X)\,\text{support}(Y)}{\text{support}(X)}$$

$$= \text{support}(Y)$$

3. Gegeben folgende Menge von Transaktionen:

TID	Items
1	D,F,G
2	C,D,E,G
3	A,D,E
4	C,D,H
5	B,C,D,F,G
6	C,F,G
7	C,D,F
8	E,G

a) Geben Sie für einen Support von 25% die vom Apriori-Algorithmus generierten Mengen C_k und L_k an.

C_k und L_k:

i. $L_1 = \{C, D, E, F, G\}$
ii. $C_2 = \{CD, CE, CF, CG, DE, DF, DG, EF, EG, FG\}$
iii. $L_2 = \{CD, CF, CG, DE, DF, DG, EG, FG\}$
iv. $C_3 = \{CDF, CDG, CFG, \cancel{DEF}, DEG, DFG\}$
v. $L_3 = \{CDF, CDG, CFG, DFG\}$
vi. $C_4 = \{CDFG\}$
vii. $L_4 = \{\}$

b) Bestimmen Sie Support und Konfidenz der Assoziationsregel $DF \to G$.

Support = 1/4, Konfidenz = 2/3

c) Nehmen Sie an, dass die Items E, F und G zum Item X generalisiert sind. Bestimmen Sie den Support von $X \rightarrow C$. Geben Sie an, ob $X \rightarrow C$ und $E \rightarrow C$ R-interessant mit $R = 2$ sind. Begründen Sie ihre Antworten.

- Support($X \rightarrow C$) = 1/2
- $X \rightarrow C$ ist R-interessant, da keine generellere Regel existiert
- $E \rightarrow C$: Support von $X = 7/8$, Support von $E = 3/8$, zu erwarten ist für $E \rightarrow C$ ein Support von $3/14$ ($= 1/2 * 3/7 = \text{support}(X, C)\frac{\text{support}(E)}{\text{support}(X)}$); tatsächlich beträgt der Support $1/8$, so dass die Regel nicht R-interessant ist.

2 TITANIC

Folgend ist der TITANIC-Algorithmus in einer Form wiedergegeben, die der Form des Apriori-Algorithmus aus der Vorlesung ähnelt.

```
1  TITANIC (I, D, minsup)
2      support({∅});
3      L_0 := {∅};
4      k := 1;
5      forall m ∈ M do {m}.p_s := ∅.s;
6      C := {{m} | m ∈ M};     // Singletons
7      loop begin
8          support(C);     // Support berechnen
9          forall X ∈ L_{k-1} do X.closure := closure(X);
10         L_k := {X ∈ C | X.s ≠ X.p_s ∧ X.s ≥ minsup};     // Pruning
11         if L_k = ∅ then exit loop;
12         k++;
13         C := Apriori-Kandidatengenerierung (L_{k-1});
14     end loop;
15     return ⋃_{i=0}^{k-1} {X.closure | X ∈ L_i};
```

1. Berechnen Sie für den Kontext auf Seite 12 der Vorlesungs-Folien (Kapitel 4, Teil 2 Begriffsverbände) die Begriffsinhalte mit einem Mindest-Support von 1 (also $33\frac{1}{3}\%$).

   $\{\emptyset, \{a, c\}, \{b, d\}, \{c\}, \{b, c, d\}\}$

2. Stellen Sie den Zusammenhang zwischen Ihren Zwischenergebnissen und den auf der Folie 12 genannten Begriffen und Formeln her.

   Die „dicken schwarzen Kästchen“ sind die minimalen Erzeuger, welche im Algorithmus in den Mengen L_k enthalten sind (in diesem Fall $\{\emptyset\}, \{\{a\}, \{b\}, \{c\}, \{d\}\}, \{\{b, c\}, \{c, d\}\}$). Die Begriffsinhalte ergeben sich als die Hüllen der minimalen Erzeuger.

3. Vergleichen Sie die einzelnen Schritte (Pruning, Kandidatengenerierung, Hüllen) des TITANIC- mit denen des Apriori-Algorithmus.

Pruning: bei Apriori Entfernung von Teilmengen, die nicht minimalen Support haben, bei TITANIC zusätzlich Entfernung nicht minimaler Generatoren

Kandidatengenerierung: Join: identisch, Pruning: bei TITANIC zusätzlich Berechnung von $X.p\_s := \min(X.p\_s, S.s)$, wobei S $(k-1)$-elementige Teilmenge von X

Hüllen: bei TITANIC zusätzlich Berechnung der Hüllen (Closure)

4. Diskutieren Sie für jeden der angegebenen Teilschritte die möglichen Performance-Gewinne und -Verluste.

Da bei TITANIC durch Entfernen nicht minimaler Generatoren (10) vor der Kandidatengenerierung (13) zunächst meist weniger Kandidaten generiert werden, kann Rechenzeit bei der Kandidatengenerierung eingespart werden. Aus dem gleichen Grund terminiert TITANIC meist nach weniger Durchläufen (11) als Apriori.