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

Dr. Andreas Hotho

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

08.05.05

Knowledge Discovery

Übungsblatt 3

Sommersemester 2005

Aufgabe 1: Stern-Schema

Die Supermarktkette IDLA möchte ihre Lagerkosten optimieren. Dazu hat sie Daten darüber gesammelt, welche Produkte in welchen Filialen und in welcher Menge über einen Zeitraum von zwei Jahren verkauft worden sind. Die Einheit der Zeitmessung sind Tage. Zur Analyse dieser Daten möchte IDLA ein OLAP System einsetzen.

a) Entwerfen sie ein Stern-Schema für die Analyse dieser Daten. Bestimmen Sie dafür vorersteinmal was die Kennzahl und was die Dimensionen dieser sind.

b) Skizzieren Sie sowohl die Kennzahlentabelle als auch die Dimensionstabellen.

c) Erweitern Sie das obige Sternschema so, daß möglichst einfach der Monatsumsatz eines bestimmten Produktes in einer bestimmten Filiale berechnet werden kann.

d) Wie würden Sie die Daten visualisieren, damit der Logistikexperte der Firma IDLA möglichst effizient die Logistikplanung für den nächsten Zeitraum vornehmen kann.

Aufgabe 2: Preprocessing und Datenbeschaffenheit

a) Welche zwei Typen des Preprocessing gibt es? Diskutieren Sie in diesem Zusammenhang den Begriff Datenverständnis?

b) In der Statistik unterscheidet man die Attributausprägungen einer vorgegebenen Menge von Daten mittels Skalen mit unterschiedlichem Skalenniveau. Nennen Sie die drei wichtigsten Skalenniveaus und beschreiben Sie sie kurz. Was bedingt ein Skalenniveau bei der Untersuchung von Daten?

c) Welches Ziel wird in Bezug auf die spätere Anwendung von Algorithmen mit dem Preprocessing der Daten im Data Mining verfolgt. Nennen Sie in diesem Zusammenhang zwei Beispiele, in denen Algorithmen bestimmte Preprocessingschritte erzwingen.

d) Nennen Sie je 5 Preprocessingschritte zu Data Cleansing und Data Manipulation und beschreiben Sie 3 davon genauer.

Aufgabe 3 Allgemeines zum Clustern

a) Beschreiben Sie kurz was man unter Clustern versteht.
b) Geben Sie verschiedene Clusterformen an.
c) Diskutieren Sie mögliche Probleme, die beim Entdecken der verschiedenen Cluster durch unterschiedliche Verfahren auftreten können.
d) Geben Sie eine typische Distanz- und eine typische Ähnlichkeitsfunktion an und diskutieren Sie die Beziehung zwischen beiden Funktionen (im allgemeinen).

Aufgabe 4 Cluster-Verfahren

Ein Kaufhaus, das seine Kunden in fünf Gruppen klassifiziert hat, möchte eine Werbekampagne durchführen. Da es zu aufwendig wäre, für jede der fünf Gruppen ein spezifisches Werbekonzept zu konzipieren, sollen sie in zwei Hauptgruppen eingeteilt werden. Dazu hat man für die Repräsentanten der einzelnen Gruppe {1,2,3,4,5} die folgenden Abstände *d* ermittelt:

D(*x*,*y*)	**1**	**2**	**3**	**4**	**5**
1	0	2	2	17	16
2	2	0	4	9	10
3	2	4	0	13	10
4	17	9	13	0	1
5	16	10	10	1	0

a) Entwerfen Sie ein Verfahren, welches ausgehend von einer Anfangsklassifikation K^0 durch den Austausch von Elementen die Klassifikation iterativ bezüglich eines Güteindexes optimiert (Austauschverfahren).
b) Ausgehend von der Anfangsklassifikation K^0={{1,2},{3,4,5}} soll mit Hilfe des Austauschverfahrens die bestmögliche Klassifikation *K* mit dem Güteindex

$$b(K) = \sum_{C \in K} \left(\frac{1}{|C|} \sum_{x,y \in C} d(x,y) \right)$$

erstellt werden.
c) Welche anderen Verfahren hätte man zur Lösung der Aufgabe auch verwenden können? (Führen Sie ein Verfahren durch, und vergleichen Sie die Ergebnisse. Zusatzaufgabe!!!)
d) Wie kann das Kaufhaus die Ergebnisse zur Aufstellung der Marketingstrategien verwenden?