Dr. Andreas Hotho, Robert Jäschke
Fachgebiet Wissensverarbeitung

11.5.2006

3. Übung „Knowledge Discovery“

Sommersemester 2006

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.

1. Entwerfen sie ein Stern-Schema für die Analyse dieser Daten. Geben sie zuerst die Kennzahl und die Dimensionen an!
2. Skizzieren Sie sowohl die Kennzahlentabelle als auch die Dimensionstabellen.
3. Erweitern Sie das obige Sternschema zu einem Schneeflockenschema unter Berücksichtigung, dass die Tagesdaten zu Monatsdaten aggregiert werden sollen.
4. Wie würden Sie die Daten visualisieren, damit der Logistikexperte der Firma IDLA möglichst effizient die Logistikplanung für den nächsten Zeitraum vornehmen kann?

2 Allgemeines zum Clustern

1. Beschreiben Sie kurz was man unter Clustern versteht.
2. Geben Sie verschiedene Clusterformen an.
3. Diskutieren Sie mögliche Probleme, die beim Entdecken der verschiedenen Cluster durch unterschiedliche Verfahren auftreten können.
4. Geben Sie eine typische Distanz- und eine typische Ähnlichkeitsfunktion an und diskutieren Sie die Beziehung zwischen beiden Funktionen (im allgemeinen).
5. Veranschaulichen Sie sich einige Distanzfunktionen, indem sie für jede Funktion in der reellen Zahlenebene ($\mathbb{R}^2$) alle Punkte mit der Distanz 1 zum Ursprung $(0, 0)$ einzeichnen.

3 Clusterverfahren

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

1. Entwerfen Sie ein Verfahren, welches ausgehend von einer Anfangsklassifikation K^0 durch den Austausch von Elementen die Klassifikation iterativ bezüglich eines Güteindex optimiert (Austauschverfahren).

2. 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.

3. Welche anderen Verfahren hätte man zur Lösung der Aufgabe auch verwenden können? (Führen Sie ein Verfahren durch und vergleichen Sie die Ergebnisse. Zusatzaufgabe.)

4. Wie kann das Kaufhaus die Ergebnisse zur Aufstellung der Marketingstrategien verwenden?

4 k-Means Clustering

1. Gegeben folgender Datensatz:

x	1	6	8	3	2	2	6	6	7	7	8	8
y	5	2	1	5	4	6	1	8	3	6	3	7

   Ermitteln Sie mit Hilfe von k-Means eine Clusterung mit $k = 3$. Verwenden Sie als Centroide die ersten drei Datentupel und verfolgen Sie die Wanderung der Centroide.

2. Betrachten Sie folgenden zweidimensionalen klassifizierten Datensatz zunächst ohne die Information über die Klasse für jedes Tupel.

x	3	3	4	4	5	6	7	7	8	9	1	2	2	3	4	5	5	6	7	7
y	1	2	2	3	3	4	4	6	5	7	3	4	5	6	6	7	8	8	8	9
Klasse	a	a	a	a	a	a	a	a	a	a	b	b	b	b	b	b	b	b	b	b

   Welches Problem ergibt sich bei der Anwendung des k-Means-Algorithmus mit $k = 2$ (d. h. zwei Clustern) auf diesem Datensatz?
   Hinweis: Überlegen Sie sich, wie das gewünschte Ergebnis aussieht. Was liefert der k-Means-Algorithmus stattdessen? (Sie brauchen das exakte Ergebnis des Algorithmus nicht auszurechnen, eine qualitative Beschreibung reicht.)