Big Data und Business Intelligence

Aufgabenblatt 02 (Verteilte DBS)

Prof. Dr.-Ing. Heiko Tapken

Wintersemester 2016/17

Testat: s. Planungen

Hochschule Osnabrück
University of Applied Sciences

Aufgabe 1. Horizontale Fragmentierung

Geben sei die Beispieldatenbank aus der Vorlesung (Abbildung 1).

EMP

ENO	ENAME	TITLE
E1	J. Doe	Elect. Eng.
E2	M. Smith	Syst. Anal.
E3	A. Lee	Mech. Eng.
E4	J. Miller	Programmer
E5	B. Casey	Sys. Anal.
E6	L. Chu	Elect. Eng.
E7	R. Davis	Mech. Eng.
E8	J. Jones	Syst. Anal.

ASG

ENO	PNO	RESP	DUR
E1	P1	Manager	12
E2	P1	Analyst	24
E2	P2	Analyst	6
E3	P3	Consultant	10
E3	P4	Engineer	48
E4	P2	Programmer	18
E5	P2	Manager	24
E6	P4	Manager	48
E7	P3	Engineer	36
E8	P3	Manager	40

PROJ

PNO	PNAME	BUDGET	LOC
P1	Instrumentation	150000	Montreal
P2	Database Development	135000	New York
P3	CAD/CAM	250000	New York
P4	Maintenance	31000	Paris

PAY

TITLE	SAL
Elect. Eng.	40000
Syst. Anal.	34000
Mech. Eng.	27000
Programmer	24000

Abbildung 1: Beispieldatenbank

a) *Horizontale Fragmentierung der Relation EMP*
Angenommen, zu den Werten des Attributes TITLE der Relation EMP ist eine alphabetische Ordnung gegeben. Gegeben seien die einfachen Attribute p1: TITLE < „Programmer“ und p2: TITLE > „Programmer“. Führen Sie eine horizontale Fragmentierung auf die Relation EMP aus, die die beiden Attribute p1 und p2 berücksichtigen.

b) *Korrektheit der Fragmentierung der Relation EMP*
Erfüllt die Fragmentierung der vorherigen Aufgabe die Regeln der Korrektheit? Begründen Sie Ihre Antwort.

c) *Eigenschaften einfacher Prädikate*
Welche Eigenschaften müssen für die Auswahl von einfachen Prädikaten gelten? Erklären Sie diese Eigenschaften. Warum sind diese Eigenschaften wichtig?

d) *Horizontale Fragmentierung der Relation ASG*

Angenommen, es gibt 2 Anwendungen, die auf Relation ASG zugreifen.

Die erste Anwendung wird von 4 verschiedenen Standorten genutzt und ist dazu da, die Dauer auszugeben, die ein Mitarbeiter zu einem Projekt zugewiesen ist. Jeder Standort ist für genau einen der Mitarbeitertypen zuständig und fragt diese ab: Anwendung an Standort 1 ist für die Manager zuständig, Standort 2 für die Consultants, Standort 3 für die Ingenieure und Standort 4 für die Programmierer.

Die zweite Anwendung wird von 2 verschiedenen Standorten genutzt. An einem Standort werden Mitarbeiter abgefragt, die weniger als 20 Stunden zu Projekten zugeordnet sind. An dem anderen Standort werden alle Mitarbeiter abgefragt, die länger zu Projekten zugeordnet sind.

Leiten Sie aufgrund dieser Informationen eine horizontale Fragmentierung ab.

e) Abgeleitete Horizontale Fragmentierung

Gegeben sei die Relation PROJ und die einfachen Prädikate p1: BUDGET < 150.000 und p2: BUDGET ≥ 100.000. Führen Sie zu diesen Prädikaten eine horizontale Fragmentierung auf PROJ aus. Führen Sie des Weiteren eine abgeleitete horizontale Fragmentierung für ASG aus. Zeigen Sie Vollständigkeit, Rekonstruierbarkeit und Disjunktheit der Fragmente.

Aufgabe 2. Vertikale Fragmentierung

Gegeben sei folgende Relation, die Teil einer Filmkritik-Datenbank ist:

```
Film (Name, Genre, Jahr, Rating)
```

Auf dieser werden folgende Anfragen ausgeführt:

```
q1: SELECT Rating FROM Film WHERE Name= 'Pulp Fiction'
q2: SELECT AVG(Rating) FROM Film WHERE Genre='Action'
q3: SELECT Name, Jahr FROM Film
q4: SELECT Name, Rating FROM Film WHERE Jahr=2011
```

a) Verwendungsmatrix
Berechnen Sie die Attribut-Verwendungs-Matrix.

b) Affinitätsmatrix
Gegeben seien folgende Zugriffshäufigkeiten für die Anfragen q_1 bis q_4 und den Knoten S_1 bis S_3:

acc1(q1) = 10	acc2(q1) = 15	acc3(q1) = 5
acc1(q2) = 0	acc2(q2) = 10	acc3(q2) = 15
acc1(q3) = 5	acc2(q3) = 0	acc3(q3) = 35
acc1(q4) = 15	acc2(q4) = 25	$acc3_{(q4)}$ = 0

Berechnen Sie aufbauend auf der Verwendungsmatrix die Attribut-Affinitäts-Matrix. Das heißt, dass Sie die Affinitäten $aff(A_i,A_j)$ zwischen zwei Attributen A_i und A_j berechnen. Die Affinität gibt jeweils an, wie häufig auf die entsprechenden beiden Attribute zusammen zugegriffen wird. Nehmen Sie weiter an, dass die Zugriffshäufigkeit immer 1 ist ($ref_i(q_k) = 1$). Berechnen Sie dabei auch die Diagonale (also auch den Fall für $A_i=A_j$), da sie diese für weitere Berechnungen benötigen werden.

c) Clustering-Algorithmus

Mit Hilfe der Affinitäts-Matrix aus Aufgabenteil b) hat man ein Maß, wie stark einzelne Attribute bzgl. ihrer Verwendung voneinander abhängen. Diese Information kann nun genutzt werden, um die einzelnen Attribute in Gruppen mit ähnlichem Verhalten zu segmentieren (Cluster zu erstellen), um die jeweiligen Gruppen mit ähnlichen Attributen letztendlich auf verschiedenen Knoten zu verteilen.

Mit Hilfe des Bond-Energy-Algorithmus‘ (BEA) lässt sich aus der Affinitäts-Matrix eine Clustered-Affinitäts-Matrix erstellen. Dazu verwendet der BEA ein globales Affinitätsmaß, das angibt, wie gut die Matrix bzgl. der Ähnlichkeit der Attribute zu seinen Nachbarn sortiert ist. Durch Permutation versucht der BEA dieses globale Affinitätsmaß zu maximieren, um dadurch möglichst ähnliche Attribute nebeneinander zu legen.

Der BEA verwendet die Stärke der Bindung (bond) zwischen zwei Attributen bei der $n \times n$-Affinitäts-Matrix, um das globale Affinitätsmaß zu bestimmen. Die Bindung zweier Attribute ist wie folgt definiert:

$$\mathrm{bond}(\mathrm{Ax},\mathrm{Ay}) = \sum_{z=1}^{n} \mathrm{aff}(\mathrm{Az},\mathrm{Ax})\ \mathrm{aff}(\mathrm{Az},\mathrm{Ay})$$

Die Bindung kann anschließend verwendet werden, um den Beitrag (contribution, cont) zu bestimmen, wenn das Attribut A_k zwischen die Attribute A_i und A_j verschoben wird:

$$cont(A_i,A_kA_j)=2\ bond(A_i,A_k)+2\ bond(A_k,A_j)-2\ bond(A_i,A_j)$$

Dabei gilt des Weiteren für die Bindung an den Seiten der Matrix, dass die Affinität null ist:

$$aff(A_0,A_j)=aff(A_i,A_0)=aff(A_{n+1},A_j)=aff(A_i,A_{n+1})=0$$

Als Beispiel zur Vorlesung würde ein Verschieben von Attribute A_4 zwischen A_1 und A_2 folgenden Beitrag machen:

$cont(A_1,A_4A_2)$=2 $bond(A_1,A_4)$+2 $bond(A_4,A_2)$−2 $bond(A_1,A_2)$

Dazu berechnen wir die Bindungen:

$bond(A_1,A_4)$=45·0+0·75+45·3+0·78=135
$bond(A_4,A_2)$=11865
$bond(A_1,A_2)$=225

und anschließend den Beitrag:

$cont(A_1,A_4A_2)$=2·135+2·11865−2·225=23550

Verwenden Sie nun dieses, um den folgenden BEA-Algorithmus anzuwenden, um die Clustered-Affinitäts-Matrix aufbauend auf der vorherigen Aufgabe zu bestimmen.

BEA Algorithm

Input: AA: attribute affinity matrix
Output: CA: clustered affinity matrix
begin
  {initialize; remember that AA is an $n \times n$ matrix}
  $CA(\bullet, 1) \leftarrow AA(\bullet, 1)$;
  $CA(\bullet, 2) \leftarrow AA(\bullet, 2)$;
  $index \leftarrow 3$;
  while $index \leq n$ **do** {choose the "best" location for attribute AA_{index}}
    for *i from 1 to* $index - 1$ *by 1* **do** calculate $cont(A_{i-1}, A_{index}, A_i)$;
    calculate $cont(A_{index-1}, A_{index}, A_{index+1})$; {boundary condition}
    $loc \leftarrow$ placement given by maximum *cont* value ;
    for *j from index to loc by* -1 **do**
      $CA(\bullet, j) \leftarrow CA(\bullet, j-1)$ {shuffle the two matrices}
    $CA(\bullet, loc) \leftarrow AA(\bullet, index)$;
    $index \leftarrow index + 1$
  order the rows according to the relative ordering of columns
end