Aufgabenblatt 3

— Indexstrukturen II —

Ausgabe am 26.05.2008
Abgabe bis 09.06.2008, 13.00 Uhr

Aufgabe 1: Eindimensionale Hashtabellen (12 P)

Sei h eine Hashfunktion, die Schlüssel auf Binärfolgen der Länge 4 abbildet. Angenommen, in einem Block können drei Records gespeichert werden. Skizzieren Sie die resultierende Organisation der Records, wenn nacheinander Records mit Hashwerten

(a) 0000, 0001, ..., 1111 eingefügt und erweiterbare Hashtabellen verwendet werden. **(3 P)**

(b) 0000, 0001, ..., 1111 eingefügt und lineare Hashtabellen verwendet werden, die eine Kapazitätsschranke von 100% haben. **(3 P)**

(c) 1111, 1110, ..., 0000 eingefügt und erweiterbare Hashtabellen verwendet werden. **(3 P)**

(d) 1111, 1110, ..., 0000 eingefügt und lineare Hashtabellen verwendet werden, die eine Kapazitätsschranke von 75% haben. **(3 P)**

Nehmen Sie jeweils an, dass die Hashtabelle initial zwei leere Blocks (0 und 1) umfasst.

Aufgabe 2: Hash-artige mehrdimensionale Indexstrukturen (10 P)

Sei $R(A, B, C)$ eine Relation, deren Extension durch

A	1	2	3	4	5	6	7	8	9	10	11	12
B	700	1500	866	866	1000	1300	1400	700	1200	750	1100	733
C	10	60	20	10	20	40	80	30	80	30	60	60

gegeben ist. Ausgehend von R soll ein zweidimensionaler Index erstellt werden, der die beiden Attribute B und C umfasst.

(a) Teilen Sie den zweidimensionalen Raum durch das Einfügen von insgesamt fünf Gridlinien so auf, dass in jedem resultierenden Bucket höchstens zwei Punkte enthalten sind. **(4 P)**

(b) Ist durch das Einfügen von nur vier Gridlinien eine Aufteilung möglich, in der jeder Bucket ebenfalls nur höchstens zwei Punkte enthält? Geben Sie entweder

eine solche Aufteilung an oder begründen Sie, warum eine solche Aufteilung nicht existiert. **(3 P)**

(c) Definieren Sie eine partitionierende Hashfunktion $h = (h_B, h_C)$, deren Wertebereich aus vier Elementen besteht, so dass jeder der vier resultierenden Buckets höchstens vier Punkte enthält. **(3 P)**

Aufgabe 3: Gridfiles **(5 P)**

Angenommen, eine Relation $R(A, B)$ ist mit einem zweidimensionalen Gridfile indiziert worden. Beide Attribute sind numerisch und können ausschließlich Werte im Bereich von 0 bis 1000 annehmen. Die Gridlinien in der A-Dimension befinden sich alle 20 Einheiten (beginnend bei 0); in der B-Dimension befinden sie sich alle 50 Einheiten (beginnend bei 0).

(a) Wie viele Buckets müssen für die Beantwortung der Bereichsanfrage

```
SELECT *
FROM R
WHERE A > 310 AND A < 400 AND B > 520 AND B < 730
```

betrachtet werden? **(2 P)**

(b) Angenommen wir wollen eine Nächste-Nachbar-Anfrage bezogen auf den Punkt $(110, 205)$ ausführen. Wir beginnen dazu die Suche im Bucket, der $(100, 200)$ als linke untere Ecke und $(120, 250)$ als rechte obere Ecke besitzt, und ermitteln in diesem Bucket $(115, 220)$ als nächstliegenden Punkt.

Welche anderen Buckets müssen noch betrachtet werden, um endgültig zu entscheiden, ob $(115, 220)$ tatsächlich am nähesten zu $(110, 205)$ liegt. Geben Sie nur solche Buckets, die zwingend betrachtet werden müssen. Geben Sie für jeden solchen Bucket die Koordinaten der linken unteren und rechten oberen Ecke an! **(3 P)**

Aufgabe 4: Baum-artige mehrdimensionale Indexstrukturen **(12 P)**

Sei $R(A, B, C)$ eine Relation, deren Extension durch

A	1	2	3	4	5	6	7	8	9	10	11	12
B	700	1500	866	866	1000	1300	1400	700	1200	750	1100	733
C	64	128	128	64	128	256	128	256	128	64	128	256

gegeben ist. Ausgehend von R sollen verschiedene zweidimensionaler Indexe erstellt werden, die jeweils die beiden Attribute B und C umfassen.

(a) Zeichnen Sie einen Index, der die beiden Suchschlüssel C und B (in dieser Reihenfolge!) enthält, wobei für jedes Attribut ein B^+-Baum verwendet werden soll.

Der Parameter n sei 2 und es wird angenommen, dass jeder Index mit der *Bulk Loading*-Methode erstellt wurde (mit maximalem Füllstand). **(4 P)**

(b) Zeichnen Sie einen k-d-Baum, wenn die Aufspaltung in der Attributreihenfolge B, C erfolgt und der Aufspaltungswert in jeder Stufe so gewählt wird, dass die Records so gleichmäßig wie möglich auf die Blätter verteilt werden. Nehmen Sie an, dass in einen Block zwei Records passen. **(4 P)**

(c) Zeichnen Sie einen Quad-Baum. Nehmen Sie dazu an, dass das Attribut B ausschließlich Werte im Bereich $[100, 1500]$ und das Attribut C ausschließlich Werte im Bereich $[0, 256]$ annimmt. **(4 P)**

Aufgabe 5: k-d-Bäume **(7 P)**

Angenommen T ist ein perfekt balancierter k-d-Baum, der eine zweidimensionalen Indexstruktur darstellt. Sei Q eine Anfrage, die nur für eines der beiden indizierten Attribute einen Wert definiert.

(a) Zeigen Sie, dass für die Beantwortung von Q insgesamt $\sqrt{n}$ viele Blätter betrachtet werden müssen. Hierbei sei n die Gesamtanzahl der Blätter von T. **(3 P)**

(b) Angenommen, T wäre alternierend in k Dimensionen aufgespalten sein ($k \in \mathbb{N}$, beliebig, aber fest). Sei Q' eine Anfrage, die für k' ($k' \leq k$) der k Attribute einen Wert definiert. Wie groß ist der Anteil der Blätter von T, die für die Beantwortung von Q' betrachtet werden müssen? **(4 P)**