Ludwig-Maximilians-Universität München
Department "Institut für Informatik"
Prof. Dr. Christian Böhm
Bianca Wackersreuther

München, 16.07.2007

Datenbanksysteme II

SS 2007

Übungsblatt 4: Relationale Anfragebearbeitung, Grundlagen der Ähnlichkeitssuche

Besprechung: 01.08.2007

Aufgabe 4-1 *Anfrageoptimierung*

Gegeben sei ein Relationenschema mit folgenden Relationen:

`Student(MatrNr, ...),Vorlesung(VorlNr, VorlTyp, ...),Dozent(DozNr, Titel, Name, ...), Hoert(MatrNr, VorlNr),Haelt(DozNr, VorlNr)`

(a) Geben Sie den kanonischen Operatorbaum für folgende Anfrage an:

"Gesucht sind alle Studenten, die bei Professor Einstein ein Seminar besuchen."

(b) Welche Optimierungsmöglichkeiten ergeben sich?

Aufgabe 4-2 *Sequentieller Scan vs. Index*

Es werde ein komplexes Selektionsprädikat auf einer Tabelle aus 1.000.000 Tupeln ausgewertet. Ein Tupel belege hierbei 100 Bytes. Für die Operation stehen 10 MByte Datenbankpuffer zur Verfügung. Die Auswertung des Prädikats benötige 10 μs CPU-Zeit. Die Daten des Plattenlaufwerkes seien wie folgt gegeben:

- $t_{seek} = 4$ ms
- $t_{lat} = 2$ ms
- Transferrate = 50 MByte/s.

(a) Zunächst sei die Tabelle in einer Datei organisiert, die sequentiell gelesen wird.

(i) Wie viel Zeit benötigt das Einlesen der Datei? Fällt die Latenzzeit ins Gewicht?

(ii) Wie viel CPU-Zeit benötigt die Selektionsanfrage?

(b) Nun sei die Tabelle in einem Index (B^+-Baum) organisiert. Eine Seite habe eine Größe von 4 KByte. Die Speicherauslastung betrage 70%.

(i) Wie viele Datenseiten werden benötigt, wenn die Datenseiten untereinander vorwärts und rückwärts verkettet sind und ein Zeiger 4 Byte benötigt?

(ii) Wie viele Directory-Seiten werden benötigt, wenn man davon ausgeht, dass der Schlüssel 20 Byte und der Zeiger auf die Sohnseite 4 Byte benötigt?

(iii) Der Index habe auf dem Selektionsprädikat optimale Selektivität (*best case*). Wie teuer ist die Auswertung (CPU und I/O)?

(iv) Wie teuer ist die Auswertung (CPU und I/O) bei schlechter Selektivität (*worst case*)? Ist die Selektion CPU- oder I/O-bound?

Aufgabe 4-3 *Äquivalenzregeln*

Beweisen oder widerlegen Sie folgende Äquivalenzen:

(a) $\sigma_{p_n \wedge p_{n-1} \wedge \ldots \wedge p_1}(R) = \sigma_{p_n}(\sigma_{p_{n-1}}(\ldots(\sigma_{p_1}(R))\ldots)$

(b) $\sigma_p(R_1 \bowtie R_2) = \sigma_p(R_1) \bowtie R_2$, falls p nur Attribute aus R_1 enthält

(c) $\Pi_l(R_1 \cap R_2) = \Pi_l(R_1) \cap \Pi_l(R_2)$

(d) $\Pi_l(R_1 \cup R_2) = \Pi_l(R_1) \cup \Pi_l(R_2)$

(e) $\Pi_l(R_1 - R_2) = \Pi_l(R_1) - \Pi_l(R_2)$

Aufgabe 4-4 *Join-Kosten*

Gegeben seien zwei Relationen R und S, die jeweils eine Größe von 10.000 Blöcken besitzen. Im folgenden soll der Join $R \bowtie S$ mittels eines Nested-Loop-Joins berechnet werden. Dabei wird als Cachestrategie Variante 3 (Skript S. 40) verwendet.

(a) Berechnen Sie die benötigte Anzahl an Plattenzugriffen bei einer Cachegröße von 1.000 Blöcken.

(b) Berechnen Sie die benötigte Cachegröße in Blöcken, um das Joinergebnis mit höchstens 100.000 Plattenzugriffen zu berechnen.

(c) Berechnen Sie die benötigte Cachegröße in Blöcken, um das Joinergebnis mit höchstens 20.000 Plattenzugriffen zu berechnen.

Aufgabe 4-5 *Recall/Precision und Sensitivität/Spezifität*

Gegeben sei das folgende gewünschte Anfrageergebnis aus einer Datenbank von 10.000 Bildern:

violine1 violine2 gitarre3 s_geige gitarre4

(a) Berechnen Sie jeweils Recall und Precision sowie Sensitivität und Spezifität für die folgenden beiden Anfrageergebnisse:

(i) Anfrageergebnis 1:

violine1 gitarre4 chmapgn s_geige gitarre3 wein01 violine2

(ii) Anfrageergebnis 2:

violine1 wein01 s_geige modell59 violine2

(b) Wie ändern sich die Werte, wenn nur jeweils die ersten k der angegebenen Ergebnisse ausgegeben worden wären ($k = 1, \ldots$)?

Aufgabe 4-6 *Distanzfunktionen*

(a) Zeigen Sie:

Für eine Metrik d gilt: $\forall o, q \in O : d(o, q) \geq 0$

(b) Sind die folgenden Distanzfunktionen für Punkte $o, q \in \mathbb{R}^n$ Metriken?

- $d_1(o, q) = \sum_{i=1}^{n}(o_i - q_i)$
- $d_2(o, q) = \max\{|o_i - q_i|, i = 1 \ldots n\}$
- $d_3(o, q) = \sum_{i=1}^{n}(o_i - q_i)^2$
- $d_4(o, q) = \sum_{i=1}^{n}|o_i - q_i|$
- $d_5(o, q) = \sqrt{\sum_{i=1}^{n-1}(o_i - q_i)^2}$
- $d_6(o, q) = \sqrt{\sum_{i=1}^{n}(o_i - q_i)^2}$
- $d_7(o, q) = \sum_{i=1}^{n-1} \begin{cases} 1, \text{ falls } o_i = q_i \\ 0, \text{ sonst} \end{cases}$
- $d_8(o, q) = \sum_{i=1}^{n-1} \begin{cases} 1, \text{ falls } o \neq q \\ 0, \text{ sonst} \end{cases}$