Klausur zur Vorlesung

Einführung in Datenbanksysteme
Datenbanken für die Bioinformatik

Annika Hinze, Zara Kanaeva

[I] Einführung in Datenbanksysteme: Bearbeitungszeit 3 Stunden (180 min) - Aufgaben 1,2,3,4
[B] Datenbanken für die Bioinformatik: Bearbeitungszeit 3 Stunden (180 min) - Aufgaben 1,2,3,5

Bitte geben Sie Ihren Namen auf JEDEM Blatt an.
Name:
Matrikelnummer:

Gewünschter Schein (nur eine Varianten ankreuzen): Bioinformatik [ ] Projektschein [ ] Übungsschein [ ]

[I] [B] 1. **Aufgabe: Entwurf** (45 min) *36 Punkte*

Betrachten Sie folgende Situationsbeschreibung:

„Patienten werden durch eine SSN-Nummer identifiziert, und sie haben einen Namen, eine Adresse und einen Geburtstag. Ärzte werden auch durch ihre SSN-Nummer identifiziert. Für jeden Arzt soll der Name und das Spezialfach gespeichert werden.

Jede Pharmafirma wird durch ihren Namen identifiziert, es wird auch die Telefonnummer gespeichert. Für jedes Arzneimittel sollen die Handelsbezeichnung (Name) und die Formel gespeichert werden. Jedes Arzneimittel wird von genau einer Pharmafirma hergestellt; die Medikamente jeder Pharmafirma sind innerhalb der Firma eindeutig durch ihren Namen identifiziert. Falls eine Pharmafirma aus der Datenbank gelöscht wird, werden auch die Informationen über ihre Produkte gelöscht.

Jede Apotheke hat einen Namen, eine Adresse und eine Telefonnumer.

Zu jedem Patienten ist sein Hausarzt gespeichert, jeder Patient hat genau einen Hausarzt. Jeder Arzt hat mindestens einen Patienten, möglicherweise mehrere. Jede Apotheke verkauft mehrere Medikamente. Ein Medikament kann von mehreren Apotheken zu einem unterschiedlichen Preis angeboten werden.

Ärzte verschreiben den Patienten Arzneimittel auf Rezept. Ein Arzt kann jedem Patienten ein oder mehrere Medikamente verschreiben. Ein Patient kann von mehreren Doktoren Medikamente verordnet bekommen. Jedes Rezept hat ein Datum und Mengenangaben bezüglich der verschriebenen Medikamente. (Rezepte haben keinen künstlichen Schlüssel!).

Pharmafirmen haben langfristige Verträge mit Apotheken. Eine Apotheke kann Verträge mit einer oder mehreren Pharmafirmen haben, eine Pharmafirma kann Verträge mit einer oder mehreren Apotheken haben. Für jeden Vertrag werden das Anfangsdatum, das Enddatum und der Text des Vertrages gespeichert. Für jeden Vertrag muss eine Person eingetragen sein, die den Vertrag betreut. "

(a) Modellieren Sie die beschriebenen Sachverhalte als ER-Schema. Fügen Sie keine künstlichen Schlüssel ein. Kennzeichnen Sie die Schlüssel der Entitäten und die auftretenden Kardinalitäten [(min,max)-Notation].

(b) Geben Sie das Relationenschema zum Entwurf an, fassen Sie die Relationen so weit wie möglich zusammen.

(c) Verständnisfrage: Wie sind Duplikat-Tupel (einer Relation) im relationalen Modell geordnet? Begründen Sie kurz.

(d) Verständnisfrage: Unter welchen Umständen ist ein Fremdschlüssel gleichzeitig Teil des Schlüssels einer Relation?

I B 2. **Aufgabe: Anfragen** (40 min) *26 Punkte*

Gegeben sei eine Datenbank mit den folgenden Relationen:

FilmRegisseur(Titel, Regisseur, Jahr)

FilmBesetzung(Titel, Schauspieler, Gehalt)

FilmKritik(Titel, Kritiker, Punktzahl)

(a) Formülieren Sie die Anfragen in SQL:

i. Gesucht: Schauspieler, die nie Regie geführt haben.

ii. Gesucht: die nach der Länge der Besetzungsliste sortierte Liste aller Filme aus dem Jahr 2000.

iii. Gesucht: Regisseure, die in einem Jahr die größte Anzahl von Filmen gedreht haben.

(b) Geben Sie die äquivalenten Ausdrücke in der relationalen Algebra und im Tupelkalkül an:

i.
```
select distinct FR.Regisseur from FilmRegisseur FR, FilmBesetzung FB
where FB.Schauspieler=FR.Regisseur and FR.Titel=FB.Titel;
```

ii.
```
select FB.Titel from FilmBesetzung FB
group by FB.Titel having count(*) > 2;
```

(c) Geben Sie einen äquivalenten Ausdruck im Tupelkalkül an:

```
select FR.Titel, FR.Regisseur, FK.Kritiker from FilmRegisseur FR,
FilmKritik FK where FR.Titel = FK.Titel
and FK.Punktzahl = select max (FK1.Punktzahl) from FilmKritik FK1
where FK.Titel=FK1.Titel;
```

(d) Verständnisfrage: Wie groß ist die Ergebnismenge des Existenzquantors ($\exists$)? Begründen Sie kurz.

I B 3. **Aufgabe: Funktionale Abhängigkeiten, Normalisierung** (40 min) *29 Punkte*

(a) Gegeben sei eine Relation R1 mit dem Relationschema R1(X,Y,Z,V,W) und mit den zwei Kandidatenschlüsseln: {X,Y,Z} und {Z,V}.

i. Geben Sie bitte die Superschlüssel an.

ii. Geben Sie bitte alle nichttrivialen funktionalen Abhängigkeiten an. Diese dürfen nicht voneinander ableitbar sein.

(b) Beweisen oder widerlegen Sie die folgenden Aussagen:

i. $X \rightarrow Y, Y \rightarrow Z \Longrightarrow X \rightarrow YZ$

ii. $X \rightarrow Y, Z \rightarrow W \Longrightarrow XZ \rightarrow YW$

iii. $XY \rightarrow Z, Z \rightarrow W \Longrightarrow X \rightarrow W$

(c) Betrachten Sie folgende Relation mit ihren funktionalen Abhängigkeiten:

R2(a,b,c,d,e,f)

$ab \rightarrow cd$

$cd \rightarrow e$

$d \rightarrow f$

i. Bestimmen Sie den Schlüssel der Relation R2.

ii. In welcher Normalform ist die gegebene Relation R2? Begründen Sie kurz.

iii. Zerlegen Sie die gegebene Relation R2 verlustlos in Relationen in BCNF, zeigen Sie die BCNF-Eigenschaften.

iv. Zeigen Sie die Verlustlosigkeit Ihrer Zerlegung aus (c).

(d) Verständnisfrage: Hat jede Menge von funktionalen Abhängigkeiten eine eindeutige minimale Hülle (minimal cover)? Begründen Sie kurz.

I 4. **Aufgabe: Physischer Entwurf, Zugriffsstrukturen** (35 min) *21 Punkte*

Betrachten Sie ein Datenbanksystem, das für die Indexierung eine Variante der B+-Bäume benutzt, bei der die Blätter die Datensätze und nicht die Verweise auf die Datensätze enthalten (Oracle: Index organized table). Es handelt sich um einen Index auf einem unique-Attribut, es kommen also keine doppelten Einträge im Index vor. Beachten Sie, dass innere Knoten einen höheren Verweisgrad haben als Datenblätter. Das Datenbanksystem soll die folgenden Eigenschaften haben:

- Blöcke haben 4096 Bytes. Der Header beträgt 96 Bytes. Die Knoten werden zu 85% gefüllt.
- Jeder Datensatz ist 300 Bytes groß. Ein Knotenzeiger (ptr) braucht 12 Bytes. Ein Datensatzzeiger (rowid) braucht 12 Bytes. Der Suchschlüssel ist 8 Bytes groß.
- Die zu indexierende Datendatei hat 1 000 000 Einträge (Datensätze).

(a) Wie groß ist der Index im worst case? Geben Sie das Ergebnis in Blockanzahl und in Bytes an.

(b) Wie groß ist der Index im worst case, wenn man in den Blättern statt Datensätzen die Verweise auf die Datensätze hat? Geben Sie das Ergebnis in Blockanzahl und in Bytes an.

(c) Betrachten Sie eine Bereichsanfrage (range query), die 10% der Daten abfragt. Die abgefragten Daten formen eine Seqenz bezüglich des Suchschlüssels, z.B. `select id from tabelle where id between x and y;`.

Welche der folgenden Zugriffsarten ist die schnellste? Bestimmen Sie zu allen vier Varianten die notwendige Anzahl an Blockzugriffen (worst case Abschätzung). Machen Sie eventuell getroffene Annahmen explizit.

i. Zugriff über B+-Baum mit Datenblättern aus (a).
ii. Zugriff über B+-Baum aus (b).
iii. Zugriff über Hash-Index.
iv. Zugriff ohne Index.

(d) Verständnisfrage: Warum können für eine Tabelle nicht zwei oder mehr Primärindexe bzw. Clusterindexe (table cluster) angelegt werden? Begründen Sie kurz.

B 5. **Aufgabe: Bioinformatikfragen** (30-40 min) *33 Punkte*

(a) Bio-Datenbanken

i. Welche Datenbank enthält alle öffentlichen DNA-Sequenzen?
ii. Welche Datenbank enthält alle öffentlichen Proteinstrukturen?
iii. Welche beiden Datenbanken enthalten öffentliche, manuelle annotierte Proteinsequenzen?
iv. Welche Datenbank enthält genetisch manifestierte Krankheitsbilder des Menschen?
v. Welche Datenbank enthält Titel, Autoren, Abstracts und teilweise auch Volltexte von fast allen wissenschaftlichen, referierten Publikationen aus den Biowissenschaften?

(b) Bioperl

i. Was ist BioPerl?
ii. Beim Zugriff von BioPerl auf EnsEMBL stehen sog. 'virtual contigs' zur Verfügung. Was sind diese 'virtual contigs'?

(c) Integration:

i. Erklären Sie den Unterschied von semantischer und syntaktischer Datenbankintegration anhand eines Beispiels!
ii. Geben Sie ein Beispiel für einen semantischen Integrationsansatz in den Biowissenschaften!
iii. Erläutern Sie das Potential und einige Probleme des semantischen Integrationsansatzes aus (b)!

(d) DAS:

i. Erklären Sie das Ziel und das Prinzip von DAS!
ii. Zu welcher Klasse der Datenbankintegrationsmethoden gehört DAS?

(e) Erklären Sie das Ziel und das Prinzip von SRS!