

# Übungen zur Vorlesung "Data Warehouses"

Sommersemester 2011

Melanie Herschel (melanie.herschel@uni-tuebingen.de)

## 5. Übungsblatt

Ausgabe: 12. Juli 2011 · Besprechung: 21. Juli 2011

**Bonusaufgabe:** Das Lösen von Aufgabe 3 ist freiwillig und bis zum 30.7.2011 möglich (das gilt nicht für die anderen Aufgaben!) Die dort erzielten Punkte werden jedoch auf Ihre Gesamtleistung angerechnet. Zur Erinnerung: zum Bestehen der Übung sind 2/3 aller Punkte regulärer Aufgaben (insgesamt 211) nötig.

### Aufgabe 1: Query Containment

(23 Punkte)

Wir betrachten eine Filmdatenbank, die folgende Relationen enthält:

Film(FID, Titel, Jahr, Rating)

FilmGenre(FID → Film, GID → Genre)

Genre(GID, Genre)

Cast(FID → Film, SID → Schauspieler, istHauptdarsteller, Rolle)

Schauspieler(SID, Name, Geburtsdatum, Geschlecht)

1. Gegeben sind folgende Datalog Anfragen. Geben Sie die jeweilige äquivalente SQL Anfrage an. Beachten Sie dabei, dass die Datalog Anfragen Mengensemantik verwenden.

$q_1(FID, T, N) : - F(FID, T, 2010, R), C(FID, SID, 'ja', R), S(SID, N, GD, 'w')$

$q_2(FID, T) : - F(FID, T, J, R), C(FID, SID, H, R), S(SID, N, GD, 'w'), J > 2000$

$q_3(N) : - S(SID, N, GD, G), C_1(FID_1, SID, 'ja', R_1), C_2(FID_2, SID, 'nein', R_2)$

$q_4(N) : - S(SID, N, GD, G), C_1(FID_1, SID, 'nein', R_1), C_2(FID_2, SID, 'ja', R_2),$   
 $F_1(FID_1, T, 2010, R), F_2(FID_2, T, 2010, R)$

2. Prüfen Sie jede Kombination  $q_i, q_j, i \neq j$  der in (1) angegebenen Anfragen auf Containment. Verwenden Sie dazu Containment Mappings (Folie 33 im Skript).

- Geben Sie für jede Kombination, für die Containment gilt, das Containment Mapping an.
- Begründen Sie für jede Kombination, für die kein Containment gilt, warum dies der Fall ist.

3. Geben Sie für jede Kombination, für die Sie in (2) ein Containment Mapping gefunden haben, an, ob auch ein erweitertes Containment Mapping (Folie 42 im Skript) existiert. Begründen Sie Ihre Antwort kurz.

## Aufgabe 2: Query Rewriting

(25 Punkte)

Wir betrachten weiterhin das in Aufgabe 1 beschriebene Filmszenario mit den dort beschriebenen Relationen. Folgende materialisierte Sichten wurden angelegt.

```
Sicht v1      SELECT DISTINCT F.FID, F.Titel, F.Jahr, F.Rating, G.Genre
              FROM Film F, FilmGenre FG, Genre G
              WHERE F.FID = FG.FID AND G.GID = FG.GID
              AND Rating > 5

Sicht v2      SELECT C.FID, Name, istHauptdarsteller, Geschlecht, COUNT(*) AS Anzahl
              FROM Cast C, Schauspieler S
              WHERE C.SID = S.SID
              GROUP BY C.FID, Name, istHauptdarsteller, Geschlecht

Sicht v3      SELECT F.FID, FG.GID, S.Geschlecht, S.Geburtsdatum, COUNT(*) AS Anzahl
              FROM Film F, FilmGenre FG, Cast C, Schauspieler S
              WHERE F.FID = FG.FID AND F.FID = C.FID AND C.SID = S.SID
              AND Geburtsdatum >= '1.1.1980' AND Rating > 8
              GROUP BY F.FID, FG.GID, S.Geschlecht
```

Folgenden Anfragen werden nun gegen das relationale Schema aus Aufgabe 1 gestellt:

```
Anfrage q1    SELECT DISTINCT Titel, Genre
              FROM Film F, FilmGenre FG, Genre G
              WHERE F.FID = FG.FID AND G.GID = FG.GID
              AND Rating > 7

Anfrage q2    SELECT DISTINCT F.FID, S.Name, istHauptdarsteller, Geschlecht
              FROM Cast C, Schauspieler S, Film F
              WHERE F.FID = C.FID AND C.SID = S.SID
              AND F.Rating > 5

Anfrage q3    SELECT F.FID, COUNT(*)
              FROM Film F, FilmGenre FG, Genre G, Cast C, Schauspieler S
              WHERE F.FID = FG.FID AND FG.GID = G.GID AND C.FID = F.FID AND C.SID = S.SID
              AND Rating > 8 AND Genre = 'Action'
              AND S.Geburtsdatum >= '1.1.1980' AND S.Geburtsdatum <= '31.12.1980'
              GROUP BY F.FID
```

1. Bearbeiten Sie pro Anfrage folgende Aufgaben. Begründen Sie dabei stets Ihre Antwort. Weisen Sie dabei insbesondere auf die in der Vorlesung besprochenen Ableitbarkeitsalgorithmen bzw. Kombinationen dieser hin.
  - (a) Von welchen materialisierten Sichten (auch mehrere sind möglich) sind oben genannte Anfragen ableitbar.
  - (b) Geben Sie jede mögliche SQL Umschreibung unter Verwendung Ihrer Ergebnisse aus (a) an.
2. Im Fall mehrerer Umschreibungen einer Anfrage, welche dieser Umschreibungen halten Sie für die sinnvollste zur Minimierung der Anfrageausführungszeit. Bitte begründen Sie Ihre Antwort kurz.

### Aufgabe 3: Data Cleaning (Bonusaufgabe)

(\*20 Punkte)

In dieser Aufgabe beschäftigen wir uns mit der Reinigung von Daten, die von <http://www.freedb.org> stammen.

1. Laden Sie sich zunächst die Daten zu ca. 10,000 CDs in der Datei `cd.csv` herunter, die auf der Veranstaltungsseite unter dem Link zu diesem Übungsblatt ebenfalls verlinkt sind. Es empfiehlt sich, diese Daten in eine relationale Tabelle zu laden. Das Schema dieser Tabelle ist in der ersten Zeile der csv Datei gegeben (alles von Typ String).
2. Identifizieren Sie in den CD Daten mindestens vier verschiedene Datenfehler (ausgenommen Duplikate). Geben Sie pro Datenfehler jeweils das fehlerhafte Tupel und eine Beschreibung des Fehlers an.
3. Ein Datenfehler, den wir in der Vorlesung kennengelernt haben, ist das Auftreten von Duplikaten. Auch solche Fehler kommen in den CD Daten vor. In der Datei `cd_gold.csv` (ebenfalls auf der Veranstaltungsseite verlinkt) finden Sie eine Liste von Duplikaten, die in dem betrachteten Datensatz vorkommen. Bearbeiten Sie folgende Aufgaben, die das Ziel verfolgen, Duplikate zu erkennen.
  - (a) Überlegen Sie sich eine Methode zur Duplikaterkennung. Diese kann z.B. der Sorted Neighborhood Methode entsprechen, oder auch einem anderen Verfahren (Vergleich aller Paare, hashbasiert, ...). Beschreiben Sie das von Ihnen gewählte Verfahren kurz. Vergessen Sie nicht, auch das Ähnlichkeitsmaß, das Sie verwenden, zu beschreiben.
  - (b) Implementieren Sie das von Ihnen gewählte Verfahren. Geben Sie dazu sowohl Ihren Quellcode als auch die Ausgabedatei, die Duplikatpaare zurückgibt, in elektronischer Form ab.
  - (c) Evaluieren Sie die Qualität Ihres Ergebnisses mittels Recall und Precision, die wie folgt definiert sind. Nehmen Sie dabei an, dass die in der Datei `cd_gold.csv` gegebenen Duplikate der Menge aller Duplikate, die im Datensatz vorkommen entspricht.

$$Recall = \frac{\text{Anzahl Duplikate, die Ihr Algorithmus korrekt identifiziert hat}}{\text{Anzahl aller Duplikate, die im Datensatz vorkommen}}$$

$$Precision = \frac{\text{Anzahl Duplikate, die Ihr Algorithmus korrekt identifiziert hat}}{\text{Anzahl aller Duplikate, die Ihr Algorithmus identifiziert hat (korrekte und falsche)}}$$

Nimmt man z.B. an, die Menge aller Duplikate im Datensatz ist  $\{A, B, C, D\}$ . Ihr Algorithmus findet Paare  $\{A, B, E\}$ . Daraus ergibt sich

$$Recall = \frac{|\{A, B\}|}{|\{A, B, C, D\}|} = \frac{1}{2}$$

$$Precision = \frac{|\{A, B\}|}{|\{A, B, E\}|} = \frac{2}{3}$$

4. In der Vorlesung haben wir verschiedene Arten von Konflikten zwischen Duplikaten kennengelernt. Soweit möglich, geben Sie jeweils ein Beispielduplikat an, das (i) Subsumption, (ii) Komplementierung und (iii) Widerspruch illustriert. Sollten Sie kein Beispiel im Datensatz finden, ändern Sie eine Dublette so ab, dass die Art Konflikt auftritt.