Database for inf3100

Igor V. Rafienko (igorr@ifi.uio.no)

Bakgrunn

I forbindelse med SQL undervisningen, skal inf3100-studenter trene på å utføre spørringer mot en enkel relasjonell database (som av historiske grunner heter inf212db). Inspirasjonen for denne testdatabasen er The Internet Movie Database (videre imdb, [1]). Kort fortalt er dette et sted der man kan finne informasjon om ca. 400 000 filmer: personer knyttet til filmene, beskrivelse av filmene, forskjellige datoer, rangeringen av filmene osv.

En god del av lærebokeksemplene bærer dessverre preg av å være kunstig enkle og/eller oppkonstruerte, noe som gjør dem relativt kjedelige å lære av. Derfor valgte vi (den undertegnede sammen med kursledelsen) å lage et utsnitt av imdb for å gi en følelse av å jobbe med virkelige data.

Dessverre viste det seg at Ifis databaseserver der inf212db opprinnelig ble lagt opp, ikke taklet relativt enkle spørringer med et fullt speil av imdb (et fullt speil har noen tabeller der antall rader er i millionklassen). Derfor måtte man ta et utsnitt av data. Jeg valgte å trekke ut de filmene som på en eller annen måte er tilknyttet til Frankrike[1].

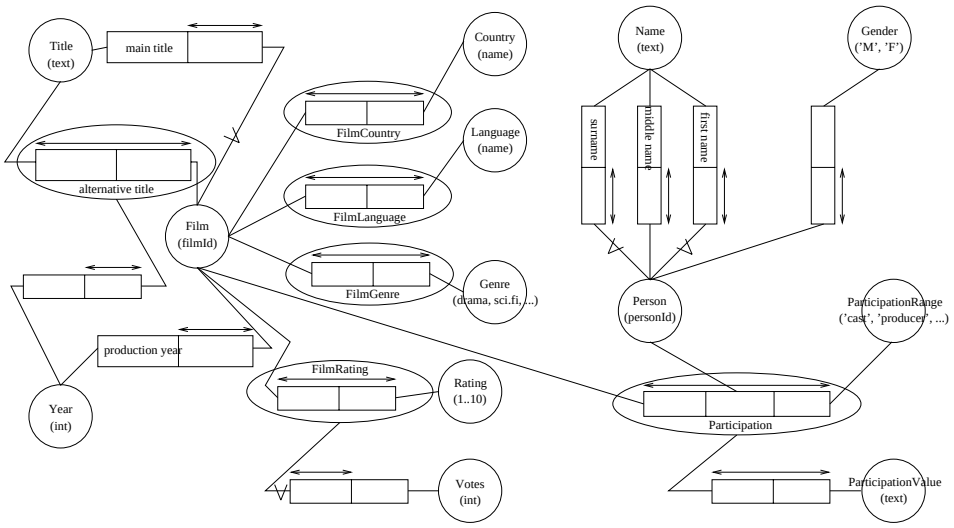
Denne første utgaven av databasen var brukt i inf212 våren 2002 og var realisert på Sybase ASE 11.9.2. Det ble etterhvert besluttet at Ifi skulle migrere til Oracle og i fjor og i år kjører databasen på Oracle 9.2i på maskinen delphinium.ifi.uio.no. DDL setninger i begge SQL varianter[2] er tilgjengelige for de interesserte. Disse inkluderer tabell-, skranke- og optimaliseringsindeksdefinisjoner.

Design

inf212db bygger rundt tre hovedbegreper – filmer, personer og deltagelse, som knytter sammen de to førstnevnte. NIAM diagrammet kan dere se i figur 1.

[1] Forøvrig betegnes slike utsnitt som ”horizontal fragmentation” i databaselitteraturen: man velger visse rader utifra tabellene basert på en rekke kriteria

[2] Oracle og Sybase har noen små forskjeller i måten man utrykker visse ting på, og databasen er ”nesten” i henhold til SQL-92 standarden



Figur 1: NIAM diagrammet for inf212db

Idet man realiserer en NIAM modell, skal man i utgangspunktet lage en tabell per begrep, men i noen tilfeller er det hensiktsmessig å fravike fra denne "regelen" og undertrykke begreper (sagt på en annen måte: et begrep blir til et attributt i tabellen som representerer et annet begrep). I inf212db ble følgende begreper undertrukket: Country, Language, Rating, Votes, Year, Title, Name, og Gender.

Vi endte til slutte med disse tabellene: AlternativeFilmTitle, Film, FilmCountry, FilmGenre, FilmLanguage, FilmRating, Genre, Participation, ParticipationRange, ParticipationValue og Person. Det er ikke alltid klart hva enkelte tabeller inneholder, og vi skal prøve å forklare de minst opplagte tilfellene.

`Rating` tabellen gjenspeiler filmens karakter i imdb. Systemet er bygd opp slik at folk kan stemme over nettet og gi hver film en karakter på en skala fra 1 til 10 ("Crossroads" ender typisk opp på 1, mens "Godfather" ender opp på den andre siden av skalaen). For hver film registreres det hvor mange personer har gitt hvilke karakterer og `Rating` tabellen inneholder nettopp disse data. Videre kan man bruke en view for å representere statistikken på en litt mer menneskevennlig måte. Viewet `AverageRating` regner ut det aritmetiske gjennomsnittet samt et vektet gjennomsnitt brukt av imdb.

Participation og ParticipationValue fortjener også en ytterligere forklaring. Participation representerer deltagelse av personer i filmer. Deltagelsen kan være av forskjellig type (alle typene er listet i ParticipationRange). For eksempel betegner

```
SQL> select * from Participation
```

```
  2  where personId = 26513 and filmId = 76914 ;
       PID   PERSONID PARTNAME                          FILMID
---------- ---------- ------------------------------ ----------
   2278671      26513 director                            76914
   2749250      26513 writing credits                     76914
SQL>
```

. . . Luc Bessons deltagelse i filmen "The Fifth Element".

pid er et kunstig innført attributt, og den brukes kun for gjøre det lettere å join'e sammen Participation med ParticipationValue (attributtet er unødvendig siden <personId,partName,filmId> danner en kandidatnøkkel. Men det er mye enklere å join'e på ett attributt istedenfor tre. Dessuten unngår man duplikasjon av informasjon).

ParticipationValue brukes for å knytte verdier til deltagelse av personer. For eksempel vil rollene som personer har spilt være listet opp i denne tabellen:

```
SQL> select * from ParticipationValue
  2  where pid = 1846849 ;

       PID           VALUE
---------- ---------------
   1846849          Leeloo
SQL>
```

. . . beskriver rollen som Milla Jovovich har spilt i "The Fifth Element"[3].

Eksempler

La oss ta et eksempel som lister opp alle roller som Milla Jovovich har spilt med tilhørende filmene:

```
SQL> select pv.value, f.mainTitle
  2  from ParticipationValue pv, Participation pa,
  3       Person p, Film f
  4  where pv.pid = pa.pid and pa.filmId = f.filmId and
  5        pa.personId = p.personId and
  6        p.surName = 'Jovovich' and
  7        p.firstName = 'Milla' ;
         VALUE                                MAINTITLE
-------------- ----------------------------------------
Mildred Harris                                   Chaplin
Lucia                                         Claim, The
Leeloo                                 Fifth Element, The
Joan of Arc     Messenger: The Story of Joan of Arc, The
SQL>
```

[3] Man kan naturligvis ikke vite at tallet 1846849 betegner nettopp hennes deltagelse. Men ved hjelp av joins med riktige tabeller kommer denne informasjonen fram.

Kildekoden

Kildekoden til databasen (både i Oracle og i Sybase) vil bli gjort tilgjengelig via kursets nettsider.

Ressurser

Siden inf212db er primært tenkt brukt i sin Oracle utgave, er det kjekt å gjøre seg kjent med Oracles SQL-dialekt. Veldig mye dokumentasjon ligger på `/local/doc/oracle`. Den mest interessante delen er sannsynligvis "SQL Reference" manualen.

Gruppelærere er også en fin ressurs. Har dere spørsmål om SQL, er de de rette personene å spørre.

Dersom dere lurer på noe *i forbindelse med inf212db*, kan dere godt spørre forfatteren direkte – `igorr@ifi.uio.no`.

Bidrag

En stor takk går til David Ranvig (davidra) for å ha konvertert imdbfiler til SQL insert setninger. Det skal sies at imdb gjør ikke akkurat sitt ytterste for å tilby sine data på et fornuftig format.

Man må også takke Rune Aske (rune) som har utsatt maskinen sin (mogwai.ifi.uio.no) for testing av imdbspeilet.

Referanser

[1] The Internet Movie Database, `http://www.imdb.com/`