Aravind Vicinthangal Prathivaathi
av4432

CSCI-620: INTRODUCTION TO BIG DATA

Assignment - 7 : Report

Note: All the Queries present here are also provided in the .sql files in the folder submitted

1. For creating the popular_movie_actors table containing movies with type “movie” and avgRating greater than 5, we just need to join the movies table and the movie_actor table and give the conditions mentioned in the question. Using this method, we create the popular_movie_actors with only the movie IDs and the actor IDs who have acted in movies with an avgRating greater than 5 and of type “movie”. The code is given in a Java file.

   SQL Query:

   ```
   Create table popular_movie_actors as select ma.actor, ma.movie
   from movie_actor as ma
   JOIN movie as m on ma.movie = m.id
   where type like 'movie' and avgrating > 5;
   ```

   The code takes 3 seconds to create the table using the SQL query.

2. This query is used to create the table L1 which indicates the first level of the lattice. Therefore, the table contains all the itemsets of size one which have a minimum support of 5. We used group by actor on the popular_movie_actors with a condition in the having clause which checks for the minimum support of 5 that is the actor must have acted in at least 5 movies.

   SQL Query for Lattice Level 1:

   ```
   Create table L1 as select pm1.actor as actor1, count(movie) as count
   from popular_movie_actors as pm1
   group by pm1.actor having count (pm1.movie) > 2;
   ```

   The query takes 86 msec to run.

3. This query is used to create the table L2 which indicates the second level of the lattice. We use the apriori-gen (join) query discussed in class. Therefore, the table contains all the itemsets of size two which have a minimum support of 5. We used group by on actor1

and actor2 on the popular_movie_actors with a condition in the having clause which checks for the minimum support of 5 that is the both the actors must have acted in at least 5 movies together.

We also do a check in the where clause of the query to make sure the actors who we use to form the itemsets are present in the previous level of the lattice, which in this case is L1.

SQL Query for Lattice Level 2:

```
Create table L2 as Select pm1.actor as actor1, pm2.actor as actor2,
count (pm1.movie) as count from popular_movie_actors as pm1,
popular_movie_actors as pm2 where pm1.actor < pm2.actor and pm1.movie =
pm2.movie and pm1.actor in (select actor1 from L1)
group by pm1.actor, pm2.actor
having count (pm1.movie)>=5
```

4. This query is used to create the table L3 which indicates the third level of the lattice. We use the apriori-gen (join) query discussed in class. Therefore, the table contains all the itemsets of size three which have a minimum support of 5. We used group by on actor1, actor2, actor3 on the popular_movie_actors with a condition in the having clause which checks for the minimum support of 5 that is all three actors must have acted in at least 5 movies together. The lattice level is created using the L2 and popular_movie_actors.

   We also do a check in the where clause of the query to make sure the actors who we use to form the itemsets are present in the previous level of the lattice, which in this case is L2.

   SQL Query for Lattice Level 3:

```
Create table L3 as Select pm1.actor as actor1, pm2.actor as actor2, pm3.actor as actor3,
count (pm1.movie)
from popular_movie_actors as pm1, popular_movie_actors as pm2,
popular_movie_actors as pm3
where pm1.actor < pm2.actor and pm2.actor < pm3.actor and pm1.movie = pm2.movie
and pm2.movie = pm3.movie and (pm1.actor, pm2.actor) in
(select actor1, actor2 from L2)
group by pm1.actor, pm2.actor, pm3.actor
having count (pm1.movie)>=5
```

5. For question 5, I have submitted a Java program that generated all the Lattice level. The program basically constructs a query for each level of the lattice and executes them. The queries are constructed from the queries used for the previous lattice levels while adding some addition restrictions required for this lattice level. The program runs until no more lattice levels can be formed, that is until an empty table is formed.

   Total number of itemset for each lattice level:

Level	No. of Item-Sets
L1	17055
L2	2462
L3	276
L4	71
L5	27
L6	5
L7	0

   Therefore, total number of lattice levels = 6

   There were 9 different actors present the last level of the lattice.
   The name of the actors in the last level of the lattice are:
   1. Bradford Hill
   2. David Gerrold
   3. Donald F. Glut
   4. G. Larry Butler
   5. George Lindsey Jr.
   6. Jason Barker
   7. Kyle Rea
   8. Robert Axelrod
   9. William Winckler

   Query used to get the actor names:

   SELECT DISTINCT m.name, m.id FROM member AS m join L6 AS l ON l.actor1 = m.id OR l.actor2 = m.id OR l.actor3 = m.id OR l.actor4 = m.id OR l.actor5 = m.id OR l.actor6 = m.id;