CSCI 620/Section 02/Mior, Introduction to Big Data, Spring 2235

Assignment 7 – Frequent itemset mining

Description

In this assignment, we are going to implement the Apriori algorithm using SQL with StackExchange data. Your code will generate a new table for each level in the lattice. Here our items will be the set of tags and our “transactions” will be the set of tags for each post. We will use a minimum support of 100.

Note: We will be ignoring the “pruning” step of the apriori-gen algorithm. This makes the code less efficient, but easier to implement, and we will have the same results.

Your tasks

1. Provide SQL to create a table L1 which contains frequent itemsets of size one. Your table should have two columns: tag1 which contains the tag ID, and count which contains the number of posts this tag was used on. Make sure you only include itemsets which meet the minimum support. **(5 points)**

2. Provide SQL to create a table L2, which contains frequent itemsets of size two (tags that were used on the same post) with columns tag1, tag2, and count (the number of times these tags were used on the same post). This should be based only on your table L1 and PostTags. This must be written as a single SQL query. **(25 points)**

3. Provide SQL to create a table L3, which contains frequent itemsets of size three with columns tag1, tag2, tag3, and count. This should be based only on your table L2 and PostTags. **(25 points)**

4. Write a program which generalizes your approach to Q3 and Q4 to generate tables for all levels of the lattice (L1, L2, …, Ln where n is the final level of the lattice). You should start with your code for Q2 which generates L1.

   You will need to programmatically generate and execute queries to create subsequent levels of the lattice (i.e. CREATE TABLE AS SELECT…). You should stop when you create an empty table at the final level of the lattice.

   Note that this is one case where you will be constructing SQL queries from strings *without* using prepared statements and parameters. This is safe since these queries are not constructed based on untrusted user input.

   (continued on next page)

Include the number of frequent itemsets in each level of the lattice in your report. For the last (non-empty) level of the lattice, include the names of the tags in each frequent itemset (by joining the table representing the final level of the lattice with the Tags table). You can perform this final query manually.
(45 points)