

C S 488/508 Introduction to Data Mining

Association rule mining algorithms

Objective

In this individual homework, you are required to get familiar with the Apriori and the FP-Growth algorithms to find association rules.

Data

Given the following transaction database

T1 = (I1, I2, I5)
T2 = (I2, I4)
T3 = (I2, I3)
T4 = (I1, I2, I4)
T5 = (I3, I4)
T6 = (I1, I3)
T7 = (I1, I2, I3, I5)
T8 = (I2, I3, I4)
T9 = (I2, I3, I5)
T10 = (I3, I5)

Q1. (50 pts) Apriori

Given minSup = 30% and conf = 70%,

- (18 pts) Show the steps of running the Apriori algorithm to get frequent 1-itemsets F_1 , candidate 2-itemsets C_2 , and frequent 2-itemsets F_2 ;
- (17 pts) From F_2 , derive all the association rules in the form of $\alpha \rightarrow \beta$ ($\alpha \neq \emptyset$ and $\beta \neq \emptyset$) that satisfy the confidence threshold;
- (15 pts) Draw the hash tree for the candidate 2-itemsets by using the hash function $x \bmod 3$ where x is the digit in an item Ix . This hash tree does not need to be a full hash tree. You just need to create ONE two-level branch and all the other branches should contain only one level.

Q2. (50 pts, CS 508 only) FP-Growth

Given minSup = 30%, please show the steps of running the FP-Growth algorithm. In particular,

- (30 pts) Show the steps of constructing the FP-tree.
- (20 pts) Derive the frequent itemsets from the SECOND item in the header table. Note that you do NOT need to show all frequent itemsets (or patterns).

Submission instructions

A zipped file `hw-lastname.zip` consisting of all the code and the PDF file.

Grading criteria

- (1) CS 508 students need to answer all the questions.
- (2) CS 488 students do not need to answer questions marked with **(CS 508 only)** although you have the freedom to work on them. Your scores will be scaled to 100. If CS 488 students answer the questions marked with **(CS 508 only)**, you will not have any points deducted if your answers are wrong; you will not get any extra points either if your answers are correct.
- (3) The score allocation has been put beside the questions.
- (4) FIVE points will be deducted if files are not submitted in the required format.