CS 412 Introduction to Data Mining

Assignment 3 Report

Haoran Yu - hyu34

1. Algorithm for Step 4 to Step 6

a. For Step 4: Apriori is adopted for this part. The algorithm works well for this problem. However, the program is somewhat slow, not because of the main algorithm but the slow function "inTrans". The function "inTrans" checked the transition by splitting the string each time and compare with each part of a query, causing some clock cycles to finish. (I would like to hire a better data structure to modify if I have some more time after I found the 15-30s running time requirement on piazza at last. I hoped there could be not too much score deducted for this issue.)

b. For Step 5: The program simply deletes the patterns which have max or closed super-patterns by one scan of the result from Step 4.

c. For Step 6: I simply follows the instruction and calculate the purity based on the result of Step 4. I first compute the $|D_t|$ and each $|D_{t,t'}|$. Then I get the

$$log\frac{f_t(p)}{|D_t|} \text{ and } log\{Max_{each\ t'}(\frac{f_t(p)-f_{t'}(p)}{|D_{t,t'}|})\}$$

which gives me the two main part for calculating the purity. I sort the result first according to the higher purity. When the patterns have the same purity, I sort them according to higher support.
For Step 6: I simply follows the instruction and calculate the purity based on the result of Step 4. I sort the result first according to the higher purity. When the patterns have the same purity, I sort them in order according to higher support.

2. Questions

a. Question to ponder A: How you choose min_sup for this task? Explain how you choose the min_sup in your report. Any reasonable choice will be fine.

   Answer:

   - ➢ I take 0.01 * the number of non-empty lines as a min_sup for each topic. 0.01 is a quite commonly used relative support. And since the size of papers in each topic is different, this method of choosing a min_sup is adopted.

b. Question to ponder B: Can you figure out which topic corresponds to which domain based on patterns you mine? Write your observations in the report.

   Answer:

   - ➢ Yes.
   - ➢ By observing the pattern files and purity files. It is quite obvious to conclude the following correspondence.
     - 0: Database(DB)
     - 1: Information Retrieval(IR)
     - 2:Data Mining(DM)
     - 3:Machine Learning(ML)
     - 4:Theory(TH)

c. Question to ponder C: Compare the result of frequent patterns, maximal patterns and closed patterns, is the result satisfying? Write down your analysis.

   Answer:

   - ➢ The result of frequent patterns is good. At least the most frequent patterns appeared in each topic matches the common knowledge of mine.
   - ➢ The maximal patterns and closed patterns result is satisfying, especially for maximal patterns. The result of maximal pattern is a lot less than the original results. However, there is almost no reduction from original patterns to closed patterns.

3. Source file description

The file trees and descriptions are as follows

```
[root] => [code.preprocessing]   => [generateDict]   => Code to generate dictionary
                                                          (Run with ./generateDict)
                                 => [tokenize]       => Codes to tokenize plain text
                                                          (Run with ./tokenize)

       => [code.partitioning]    => [lda-c-dist]     => Codes to assign topics(given)
                                 => [reorganize]     => Codes to re-organize the terms
                                                          (Run with ./reorganize)
                                 => [result]         => Result for lda-c-dist

       => [code.freqPatternMining]=> Code for mining frequent patterns for each topic
                                   (Run ./fpmining #topic #minsup. Exp:./fpmining 0 0.01)

       => [code.cloMaxPattern]   => Code for mining maximal/closed patterns
                                    (Run with ./closmining #topic. Exp:./closmining 0)
                                    (Run with ./maxmining #topic. Exp:./maxmining 0)

       => [code.rerankByPurity]  => Code for re-ranking by purity of patterns
                                    (Run with ./reranking #topic. Exp:./reranking 0)

       => [patterns]             => Result: pattern-0.txt~pattern-4.txt

       => [max]                  => Result: max-0.txt~max-4.txt

       => [closed]               => Result: closed-0.txt~closed-4.txt

       => [purity]               => Result: purity0.txt~purity-4.txt

       => Result: title.txt

       => Result: topic-0.txt~topic-4.txt
```