CPE403-CSC403 Advanced Data Management Techniques
(Data Mining)

Assignment 2 – Construction of FP-tree

Total: 20 Marks

Due Date: 4:59:59 p.m., 13 April, 2012

Reminders

- You are NOT allowed to COPY code/report from Internet or other groups. Any plagiarism cases will be seriously punished!
- You need to implement the code for the algorithm by yourselves. Simply calling existing functions/procedures for the algorithm is NOT allowed.
- For late submission, a penalty of 2 marks per day (including Saturdays, Sundays, and public holidays) will be applied after the deadline.
- Programming language: C++ or Java only
- Operating System Platform: Windows or Linux only

Implement and analyze the data structure FP-tree. Reference to FP-tree can be found in the lecture notes and the following research paper:

- J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. ACM SIGMOD 2000.

The input & output of your program should be as follows.

Input:

- Transaction file name
- Number of transaction
- Number of items
- Maximal length of transactions
- Support threshold

The input transaction file can be downloaded from **http://fimi.ua.ac.be/data/**

Each file has the following format:

1. Each line in the file is a transaction.
2. A transaction is recorded in form of "item_1 item_2 ...", where each item is represented as an integer.

Output:

- A file that prints the FP-tree in depth-first order, where each node is printed together with its frequency associated with the node
- Number of nodes in the FP-tree
- Number of leaves in the FP-tree
- Height of the FP-tree
- The minimum, average, and maximum fanout (number of children of a non-leaf node) of the FP-tree

What to submit:

1. Source code of your implementation (no need to print the code in your report)
    - Well-commented code
    - Include Makefiles if necessary
    - Remove the binary executable program if any
2. A README file. Please name it **README.txt** This file should include three sections:
    - Your group ID and group member names
    - Language used and instructions on how to run your programs.
3. A report. Your report should address at least the following issues.

    a. Indicate what other (softcopy) you have submitted in addition to this report.
    b. Description of your implementation, including key techniques. Please analyze what are the smartness and efficient ways in your implementation.
    c. What are the most expensive operations in your implementation of FP-tree construction?
    d. According to your experience, what in FP-tree could be improved?
    e. Report the number of nodes and leaves, the height, and the minimum, average, and maximum fanout (number of children of a non-leaf node) of the FP-tree, the runtime and the memory consumption of your program, by setting the minimum support threshold to 50%, 10%, 5%, and 1%, respectively, for each of the datasets you test.

Submission Instructions

Please package all of your files (including the code, the README.txt file, and your report Asg2_Report_GroupXXX.pdf) into a ZIP file, name it as “Asg2_GroupXX.zip”, where XX is your group ID.

Submit the package file with the Subject “**ASG2 SUBMISSION: GROUP XX**” to the following email: **cpecsc403@gmail.com**

You ALSO need to submit a hardcopy of your repot to my Office: N4-2B-43 by hand or put it in my pigeon hole in the general office by **4:59:59 p.m., 13 April, 2012**