Homework 1

1 Algorithms Explanation

1.1 Step 4

In this step, I choose Apriori as the Frequent Pattern Mining algorithm. There are two main steps: First step is finding the candidate set by calculating the occurrence of patterns started from length one. Second step is choosing the min support (Here I choose 1% of the number of total lines in the topic.) and delete the patterns having lower support than the min support to get the frequent set. These two steps are repeated until the new frequent set only have one or zero patterns. Frequent sets and corresponding support got in each iteration are saved and written in the pattern flies.

1.2 Step 5

In this steps, closed and max patterns are mined according to the definition. Closed pattern means there is no super set of this pattern that has the same support. So I traverse each pattern in the pattern file to check whether it has a super set with the same support. Max pattern means there is no super set of this pattern that is frequent pattern. Since all the patterns in the pattern files are frequent, I traverse each pattern to check whether it has a super set in the same file.

1.3 Step 6

In this step, I use the given equation to calculate the purity. Here, $f(t,p)$ is the support of the pattern p in topic t; $f(t',p)$ is the support of the pattern p in topic t'; $D(t)$ is the number of lines in topic t; $D(t,t')$ is the number of unique lines in the combination of topic t and topic t'. For combination of support and purity, I use the equation as: $log_2(support) + purity$. This is because purity is too small, if directly combined with support, the rank can hardly change.

2 Question Answer

2.1 Ponder A

I choose 1% of the number of total lines in each topic file as min support. I think it is reasonable. For example, in 10,000 titles, a pattern appears more than 100 times is relatively frequent.

2.2 Ponder B

Topic 0: Database
Topic 1: Data Mining
Topic 2: Artificial Intelligence
Topic 3: Computer System
Topic 4: Information Retrieval

2.3 Ponder C

Compared with frequent patterns, closed patterns just removes several patterns that contains same terms. These two type of patterns can just help me make a rough guess of the topic. But max patterns contains much fewer patterns than other two files. Through observing the patterns in max pattern files, I can basically determined the topics.

3 File List

Preprocessing.py: Step 1 - Step 2 Partitioning.py: Step 3 FPMining.py: Step 4 MaxCloseMing.py: Step 5 Purity.py: Step 6 WritePharse.py: Map id to terms