Declarative Itemset Mining

Exercise 1

In this exercise, you will work with the following tools :

— CHOCO-MINING : A Java library designed for solving itemset mining problems, built on the CHOCO-SOLVER framework.

— The SPMF library : An open-source Java-based software and data mining library specializing in pattern mining (SPMF).

Question 1 • Clone the GitHub repository of CHOCO-MINING (link).

Question 2 • Open the file `ExampleClosedItemsetMining.java` and perform the following tasks :

1. Review the code in detail.
2. Run the main method.
3. Run it on other datasets such as *mushroom* or *chess*.
4. Display the number of resulting patterns.
5. Display the execution time.

Question 3 • Add the frequency constraint : $freq(P) \geq \alpha$.

Question 4 • Add a constraint on the size of the returned patterns : $size(P) \geq lb$.

Question 5 • Now, replicate the tasks using SPMF. Run the `.jar` file available in your local repository. The goal is to run LCM for closed itemset enumeration, relaunch with different thresholds for frequency, and also for pattern size.

Question 6 • Add a constraint, called `CategoryConstraint`, to the file 'ExampleClosedItemsetMining.java' to model the following problem : Consider a dataset with n items, organized into categories of size $catSize$ (e.g., household products, appliances, etc.). The dataset is divided into $nbCat = n/catSize$ categories, with items that do not belong to any category (but do not exceed the size of $catSize$). Figure 1 shows an example with 8 items, 2 categories of size 3, and 2 items that do not belong to any category. The task is to create a constraint model that enumerates all closed itemsets composed of items belonging to at least m categories :

$$\texttt{CategoryConstraint}(P) \equiv \sum_{i=1}^{nbCat} \prod_{j=1}^{catSize} P_i \geq m$$

For example, in the dataset shown in Figure 1, with $m = 2$, the following pattern is produced : BEF.

Question 7 • How can this `CategoryConstraint` be taken into account in SPMF ?

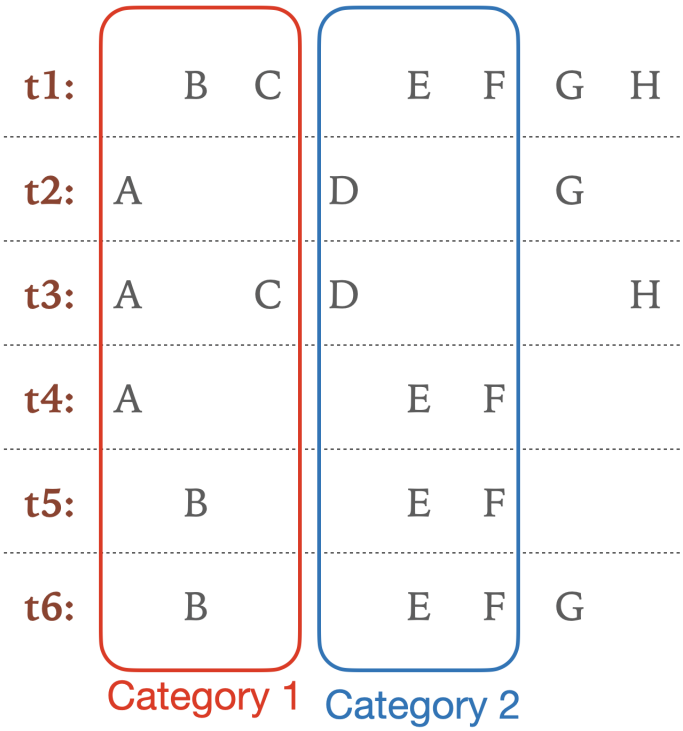

FIGURE 1 – Items categories illustration.