Ludwig-Maximilians-Universität München
Institut für Informatik
Dr. Eirini Ntoutsi
Erich Schubert

Knowledge Discovery in Databases

SS 2012

Übungsblatt 2: Association rules & Frequent Itemsets Mining

Aufgabe 2-1 *General understanding part* (Uses of FIM and association rule mining)

Frequent itemsets and associations rules mining are not restricted to market basket analysis, but rather they can be applied in several other settings. Below are given some examples of problems / applications. How would you model them for frequent itemsets / association rules mining? What are the transactions and the items in each case? What kind of rules / frequent itemsets do you expect to find?

- Online shop e.g. Amazon
- Fast food
- University library
- Technological forum

Aufgabe 2-2 *Algorithmic part* (Apriori-Algorithmus)

Given a set of items I = {A, B, C, D, E, F, G, H, I, K, L, M} and a set of transactions T according to the following table:

Set of transactions T

Transaction ID	items in basket
1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

(a) Determine the frequent item sets for a minimum support of 30%, using the Apriori algorithm from the lecture. Explicitely give the candidate sets after the join and prune steps each as long with the actual frequent items and their support.

(b) Determine all association rules that can be derived from the frequent item set $\{BEGH\}$ with a confidence of at least 60% and a support of 4 (frequency 30%). Use the monotonicity as introduced in the lecture.

(c) Considering the same minSupport settings, what are the closed frequent itemsets (CFI)? How do they relate to the FI set?

(d) Considering the same minSupport settings, what are the maximal frequent itemsets (MFI)? How do they relate to the FI, CFI sets?

Aufgabe 2-3 *Project part* **(Stack overflow dataset)**

The goal of this exercise is to apply the frequent itemsets and association rules mining concepts to a real application scenario. We will come across some unexpected challenges even before applying the Apriori algorithm.

The data set we are going to work with is a dump from the web site `http://stackoverflow.com/`, which is a a language-independent collaboratively edited question and answer site for programmers. Users posts their questions and assign them to some predefined category like “java”, “php”, “mysql”; each question can be assigned to maximum 5 categories. Our goal is to find what are the most common categories combinations in the questions.

In the computer tutorial session (to be announced) we will demonstrate all the steps of the KDD process from data selection to data cleaning, data transformation, pattern extraction and pattern evaluation. You can follow all or some of these steps and get hands-on experience.

It is commonly stated that “data pre-processing might take more than half of the total time spent for solving the data mining problem”. What is your opinion on this, based on the demonstration?