TDT4300 - Assignment 2

Håkon Ødegård Løvdal and Aleksander Skraastad

February 2015

Contents

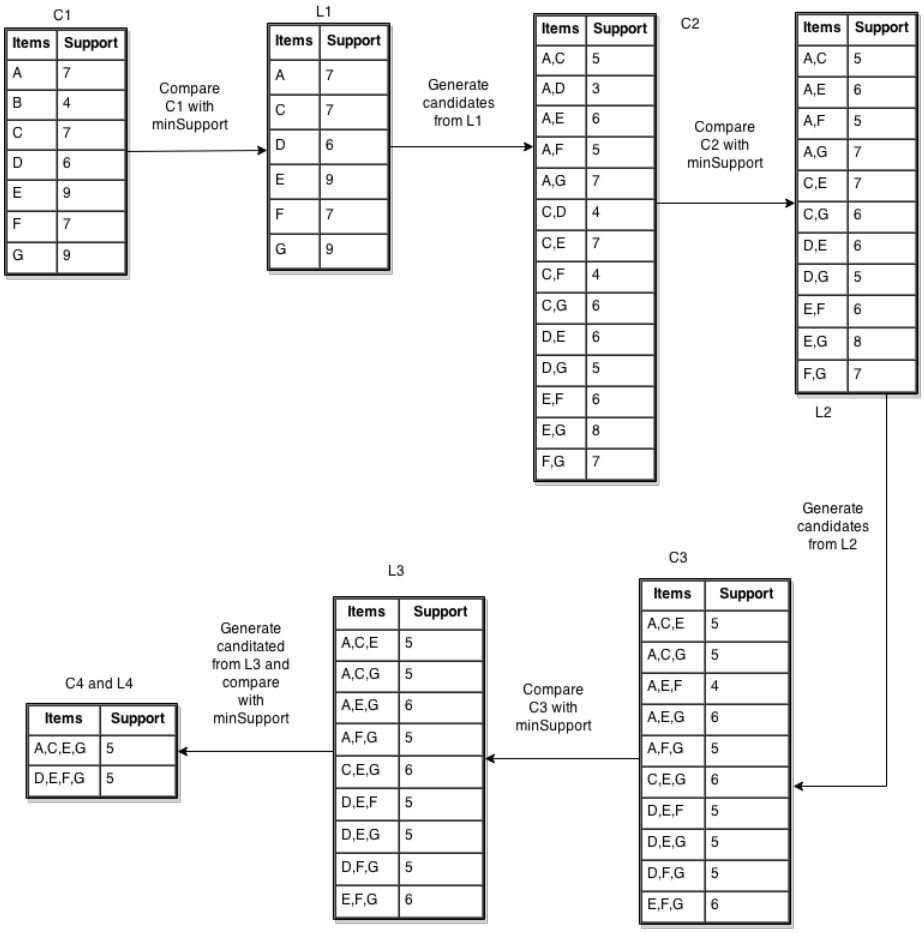


Figure 1: Finding frequent itemsets

1 Task 1

We generated all assosiation rules for the given data by applying the Apriori principle (with $F_{k-1} \times F_{k-1}$ algorithm for candidate generation. The process is illustrated in figure 1. The main concept is the following:

First calculate the support count for every item. The task asks us to consider the support threshold $minsup = 0.5$, wich means that the support count must by ≥ 5. This gives us the L1. From L1 we generate every possible itemset, and calculate their support count. The process is similar for every iteration of the algorithm until we have generated the largest frequent itemset. Notice that from L2 to C3 the set {A,C,F} is excluded/pruned because of the fact that {C,F} is not in the frequent itemsets.

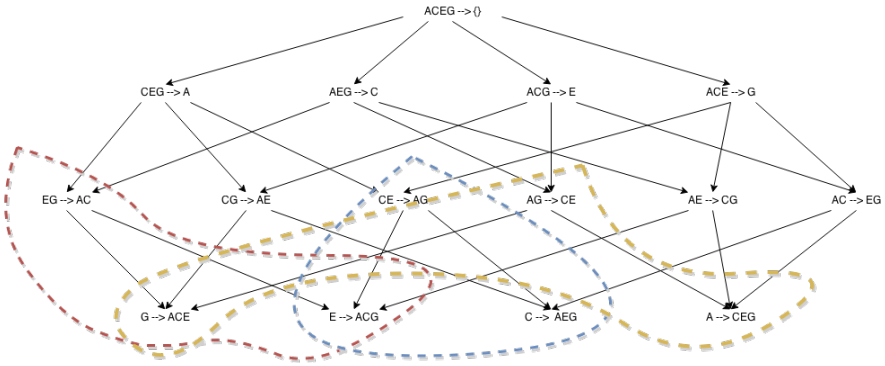


Figure 2: Example of rule generation

In figure 2 we show an example of rule generation, and how some subsets are pruned because of their confidence being < 0.8, wich are the provided $minconf$. The procedure is the same for the subset {D,E,F,G}

Following this procedure we get the following set of association rules (all wich support the minconf, but not all are strong):

Rule	Support	Confidence
ACE $\rightarrow$ G	5	5/5
ACG $\rightarrow$ E	5	5/5
AEG $\rightarrow$ C	5	5/6
CEG $\rightarrow$ A	5	5/6
AC $\rightarrow$ EG	5	5/5
AE $\rightarrow$ CG	5	5/6
CG $\rightarrow$ AE	5	5/6
D $\rightarrow$ EFG	5	5/6
DEF $\rightarrow$ G	5	5/5
DEG $\rightarrow$ F	5	5/5
DFG $\rightarrow$ E	5	5/5
EFG $\rightarrow$ D	5	5/6
DE $\rightarrow$ FG	5	5/6
DF $\rightarrow$ EG	5	5/5
DG $\rightarrow$ EF	5	5/5
EF $\rightarrow$ DG	5	5/6

2 Task 2

Please see the attached .jar file.

Usage:

```
java -jar TDT4300_assignment2.jar <dataSetSize> <minSup> <minConf>
```

Example:

```
java -jar TDT4300_assignment2.jar small 0.5 0.8
```

3 Task 3

3.1 Task 3.1

The task here is to calculate the Euclidean and Manhattan distance matrices between the following coordinates:

$Porto = [0, 2]$
$Oslo = [3, 6]$
$Moscow = [7, 6]$
$Munich = [3, 3]$

Euclidean distance:

$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$

Manhattan distance:

$d = |x_1 - x_2| + |y_1 - y_2|$

3.1.1 a) Euclidean

	Porto	Oslo	Moscow	Munic
Porto	0	5	8.0623	3.1623
Oslo	5	0	4	3
Moscow	8.0623	4	0	5
Munich	3.1623	3	5	0

3.1.2 b) Manhattan

	Porto	Oslo	Moscow	Munic
Porto	0	7	11	4
Oslo	7	0	4	3
Moscow	11	4	0	7
Munich	4	3	7	0

3.2 Task 3.2

The task is to calculate the Jaccard Coefficient to illustrate the similarity between two subsets of countries from the map.

Yellow = y
White = w

$y = [\prime norway\prime, \prime sweden\prime, \prime finland\prime, \prime russia\prime, \prime estonia\prime, \prime latvia\prime, \prime lithuania\prime, \prime belarus\prime, \prime poland\prime, \prime germany\prime, \prime netherlands\prime, \prime denmark\prime]$

$w = [\prime england\prime, \prime france\prime, \prime belgium\prime, \prime netherlands\prime, \prime denmark\prime, \prime switzerland\prime, \prime sweden\prime, \prime austria\prime, \prime hungary\prime, \prime romania\prime, \prime czechrepublic\prime, \prime slovakia\prime, \prime ukraine\prime, \prime belarus\prime, \prime lithuania\prime, \prime germany\prime, \prime latvia\prime, \prime poland\prime]$

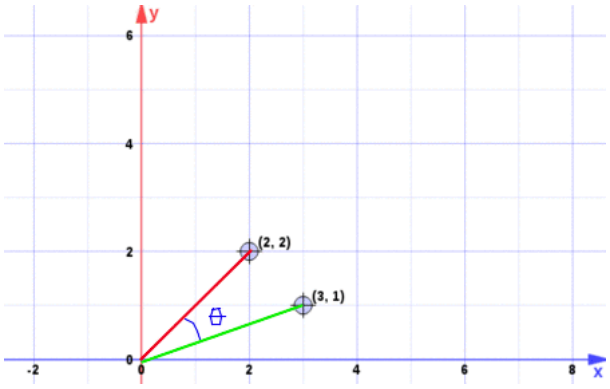
Using the following formula for the Jaccard Coefficient we can calculate the similarity:

$\frac{M_{11}}{M_{01}+M_{10}+M_{11}}$

$\frac{8}{4+10+8} = 0.3636364$

3.3 Task 3.3

Tasks a) and b):



3.4 Task 3.4

Cosine similarity 2.

$a = [84, 21, 0, 42]$
$b = [4, 1, 0, 2]$
$x = [2, 2, 0, 3]$

Given these document vectors, the cosine similarity between vectors $sim(a, x)$

and $sim(b, x)$ were as follows:

$$sim(a, x) = \frac{84*2+21*2+0*2+42*2}{\sqrt{84^2+21^2+0^2+42^2}*\sqrt{2^2+2^2+0^2+3^2}} = 0.84680979844$$

$$sim(b, x) = \frac{4*2+1*2+0*2+2*2}{\sqrt{4^2+1^2+0^2+2^2}*\sqrt{2^2+2^2+0^2+3^2}} = 0.84680979844$$

As we can see, the vectors a and b appear quite different at first glance. However, if we examine the relative ratios between each axis in both of the vectors a and b, we see that they share the same coeffitients. This means that even though the length of the vectors are different, they are still parallell, thus yielding the same angular difference from vector x.