Data Mining - Report 2

Damage

1 Problem and Approaches

The task was to determine patterns in the choice of courses computer science students of the university of Helsinki take. Therefore we were given a data set containing the courses and their metadata taken by students over some years.The methods of representation we used in a previous attempt, frequent itemsets, was not optimal. It did solve our problem, but it also represented uninteresting and redundant information. Therefore, we were introduced to the representation methods of maximum and closed itemsets during class.
Our approaches solving the problem this week were therefore alternated. We extended our own implementation for itemset generation to also include the new representation methods maximal and closed frequent itemset. Furthermore we concentrated on the second part of the association rule mining stategy, the actual rule generation. In our implementation, we also considered different measurement methods to determine the value of a rule and an itemset.
The eclat software was still used, but mainly for the purpose of verifying the outcome of our own implementation.

2 Data

In addition to the course information from last week, we now had a file which contained more specific data about the courses. This metadata consisted of the time period the course was given, the term, its level and copulsoritiy. Furthermore, it contained information about subprogramms, which it might belong to.

This missing information caused insufficiencies in the interpretation of last weeks results, therefore we are now able to analyse the data more intensly. However, integrating this information caused problems as well, which are further discussed in the next section.

3 Transformation and Command line arguments

Both the meta data in course_details.txt and actual FID data in course_num.txt were transformed. Firstly, only courses with meta data information were taken into account and FID's not present in course_details were omitted when reading course_num. Secondly all the courses were grouped to single entity by FID.

Each so acquired course instance then had following attributes:

1. FID - fid
2. NAME - course name, lower case and slugified
3. YEAR - sequence of years the course has been taught, i.e. [1999, 2000, 2004]
4. SUBPROGRAM - subprogram of the course
5. COMPULSORY - P:yes V:no ?:not known

The code and semester information were omitted because they were thought to be non- relevant. When the data was transformed it was easy to only take

into account for example courses that are compulsory, taught on certain year interval, etc. We also added some command line tools to restrict the courses.

For example:
> python prob2.py t=0.4 c=0.2 year=2006-2011 compulsory=V strip=2

would only look at the non-compulsory courses that were taught in years from 2006 to 2011 and after it would strip of all the transactions with 2 or less items. In this case minimum support would be 0.4 and minimum confidence 0.2. The reason to apply these kind of restrictions is to look only subset of courses which might give more interesting results.

As mentione above, we removed course without metadata from our dataset. However, we had considerd different ways of handeling those courses. Each of the three possibility had some disadvantages which would affect our results and interpretation.

A first possibility was to leave those course with unknown values for the metadata. However, this would have leaded to futher complications as soon as the metadata was used to limit the considered data to certain courses.

A secound possiblity was to just skip those course and exclude them form the data. Yet, this would affect statisical data such as the average amount of course taken by a student, but also the frequency of the other courses. Furthermore, if a transaction only consits of courses with unkown metadata, the whole transaction would be missing.

A third possiblity was to remove the whole transaction, if one of its courses is without metadata. This would not affect basic statics, such as average number of courses, assuming that courses without metadata are distributed evenly, which we cannot garantie. Nevertheless, we would minimize our dataset of transactions, might as well remove patterns.

After some discussions, we decided on the secound option, removing the courses form our dataset. The disadvantages seemed to be the easierst to handle, so it was the least evil.

n

4 Implementation

5 Results and Conclusion

6 Teamwork Evaluation