Assignment - 3

Association Rule Mining

Deadline - 10th October 2022 11:55 PM

Necessities -

- **Allowed Languages**: *C, C++, and Python.* Note that whichever language you choose, you are not allowed to use any specific libraries for the direct implementation of the algorithms; however, standard libraries are allowed.
- All the tasks are compulsory.

Objective: Build an association rule-based movie recommender system

Input: A data set with **100836 ratings** and **3683 tag** applications across **9742 movies** is given. It describes a 5-star rating activity from MovieLens, a movie recommendation service. Refer to README for more details.

All ratings are contained in the file **ratings.csv**. Each line of this file after the header row represents one rating of one movie by one user and has the following format: **userId, movie, rating, and timestamp.**

Movie information is contained in the file **movies.csv**. Each line of this file after the header row represents one movie.

It has the following format: **movieId, title, genres**

Dataset Link: https://files.grouplens.org/datasets/movielens/ml-latest-small.zip

Data Preprocessing:

1. Form the transactional data set, which consists of entries of the form <user id, {movies rated above 2}>.
2. Consider only those users who have rated more than 10 movies.
3. Divide the data set into 80% training set and 20% test set. Remove 20% of movies watched from each user and create a test set using the removed movies. The remaining 80% of data is training set.

Tasks:

1. From the training set, extract the set of all association rules of form X→Y, where X contains a single movie and Y contains the set of movies from the training set by employing the apriori or FPgrowth approach and set **minsup= 0.01 and minconf=0.1.**

2. **Recommendation**: Create two lists: The first list consists of top-100 association rules sorted based on the support. The second list consists of top-100 rules based on confidence. Select the rules which are common to both lists. Sort the common rules based on confidence.
3. For each user in the test set, select association rules of the form X→Y, where X is the movie in the training set. Compute the average precision and average recall by varying the number of rules from 1 to 10 and plot the graphs.
   For example, consider rule X→Y, where X is the movie from the training set. The set of movies in set Y is recommended movies. In this manner, if we consider N rules, combining the movies on the right side of each N rule constitutes the set of recommendations, say R. The intersection of R with the test set is called the **hit set**. The ratio of the hit set and test set is equal to **recall**. The ratio of the hit set and recommendation set is equal to the **precision**.
4. Take a sample example of users and their movie ratings from the test set and Display precision and recall graphs.
5. Justify the selection of the algorithm you choose in the report. Explain clearly how you have built the recommender system and what optimizations you have used in the algorithm (you can also use specific optimization techniques for this application).

Submission Instructions-

- Naming convention : <TeamName>_Assignment3.zip . Submit a **zip file** containing your code files and the report pdf for the tasks mentioned in the assignment.
- There should be a readme file, 3 text files for results, 2 code files, and the report in the zip file.
    - The readme file must contain information about how to run the code and the file structure in the zip file. It should be named as <TeamName>_readme.MD.
    - The codes should named as <TeamName>_ruleminer.cpp and <TeamName>_recommender.cpp (or c or py extension as required).
    - The results file should be named as <TeamName>_RulesMaxSupport.txt, <TeamName>_RulesMaxConf.txt, <TeamName>_AssocRules.txt
    - The report name should be <TeamName>_report.pdf.
- Submission from one member is sufficient.
- Please refrain from any kind of plagiarism; else, the cases involved will be dealt with strict penalties.
- **Any submissions adhering not to the defined format will incur the penalty**