Binghamton University
Computer Science Department
CS435/535 Fall 2020
Semester Long Project
Assigned on 26 August 2020
Final Due on 15 November 2020

Total Points: 100

This is a semester-long project, encouraging students to dive into the research in data mining. Specifically, this project concerns with the research on developing a recommender system. You are asked to conduct the research on recommender systems independently by yourself. No collaboration is allowed.

Recommender systems have been studied extensively in the literature and have also been found popularly useful in many real-world applications. A recommender system is considered in the scenario where we have m different users and n different items in a specific application (e.g., in an E-commerce application we have n different customers potentially intending to buy m different commercial items) where a user i may give an item j a rating value k based on this user's preference ($i \in [1,\dots,m]$; $j \in [1,\dots,n]$; $k \in [1,\dots,K]$). Essentially, the whole user preference rating data can be represented as a matrix of n by m where each element of this matrix is the value k if the user in question gave a rating for the item in question, and 0 if no such rating was given yet. Initially, all the elements of this matrix are 0. After we have collected a certain amount of such user preference rating data, this matrix becomes partially filled out with different non-zero values. The goal of a recommender system is that, after we have collected a certain amount of the user preference rating data, the system is able to predict the rating value k of user i for item j if user i has not yet given such a rating. In other words, given such a partially filled out matrix, a recommender system shall be able to fill out all the predicted rating values for those elements with the current value of 0. That is, a recommender system mathematically is a solution to a matrix value completion problem.

This semester-long project consists of three phases. Undergraduate students are required to complete the first two phases only and graduate students are required to complete all the three phases.

In the first phase, you dive into the literature on recommender systems and develop a solution to this problem by either proposing a new recommender system by yourself or identifying an existing, working recommender system from the literature. You then implement this system using whatever language and/or environment you are comfortable with.

In the second phase, you use the given dataset (file: train.txt) to train and evaluate your solution. Note that you may want to split the given dataset into two parts with one for actual training and the other for evaluation. The given dataset (file: train.txt) is a partially filled out rating data matrix. The matrix is given in an ASCII text file where each row is a non-zero element of this matrix with three entries: the user ID number, the item ID number, and the rating value. The rating value is in the range of [1, 5] in integer where 1 means that the user likes the item the least and 5 means that the user likes the item the best. The matrix has 943 users (m = 943) and 1682 items (n = 1682). Your goal is to fill out all the entries of this matrix with the current value as 0 (or no value yet). When you report your result at the end of this phase, you must make sure that your code generates an ASCII text file for the whole rating matrix with each element filled out (either as given or as the predicted rating value by your code). The output text file must follow the format of one element of the matrix as one row beginning with the element for the first user for the first item, followed by the first user for the second item, …, the first user for the last item, the second user for the first item, …, the second user for the last item, …, the last user for the first item, …, the last user for the last item; each row of the file has three numbers: user ID number, item ID number, and the rating value, separated by a space. A sample file is given as submit_sample.txt. **Note that we use a script to grade your turned-in file and any violation of this format requirement shall result in the penalty of no credit for this part.**

In the third phase, you write up a mini-paper of no more than three pages (12 point font, single spacing, one inch margin) answering the following questions concisely.

1. Briefly describe how your solution works in English (not in pseudo code).
2. Briefly describe your strategy for the cold start issue (i.e., predicting an element in the matrix that relates to a new user who never gives any preference rating before or to a new item that no user ever gives rating to this item before).
3. Give two other examples of using a solution to the matrix completion problem in the real-world other than recommender systems, and be specific in the physical meanings of the two dimensions as well as the entries of the matrix.
4. Read at least three papers on recommender systems published within last three years other than the one you implemented in this project if you elected to take an existing solution from the literature, and write up a mini-survey for the papers you have read.
5. Based on the literature you have read as well as your experience in this project, discuss what the open issues (other than cold start) are on the topic of recommender systems and how people are attempting to address these issues or these issues are purely open with no one ever attempted to address them.

When and what you need to turn in: By noon of the due date, you must submit a zipped package containing the source code of your implementation of the recommender system with appropriate comments and documentations in the code, a README file to explain how to compile and run your code under what specific environment, and a text file containing the output matrix following exactly the format requirement stated above. For graduate students, you also need to turn in the mini-paper with requirement stated above.

Points Distribution: 65 points are given to Phase 2 grading where 60 points are given to the accuracy grading and 5 points are given to the documentations for the source code, and 35 points are given to Phase 3 grading where the breakouts are 5, 5, 6, 9, 10 to the five questions, respectively.

Bibliographical notes: In the early literature on recommender systems, different techniques of collaborative filtering are extensively used to develop the recommender systems [1,2]. [3] is a special issue focusing on the literature on this topic. One of the challenging issues in the literature on recommender systems is that it is difficult to give an appropriate prediction for a user who does not leave any history prediction data in the matrix (i.e., a "new" user), or for an item that does not have any history prediction data in the matrix (i.e., a "new" item). This is called a *cold start* issue in the literature.

A final note: you must complete this project independently and no collaboration is allowed.

References

[1] J. Breese, D.. Heckerman, and C. Kadie, Empirical analysis of predictive algorithms for collaborative filtering, *Proc. Conf. Uncertainty in Artificial Intelligence,* (UAI98) 1998

[2] J.L. Herlocker, J.A. Konstan, J.R.A. Borchers, and J. Riedl, An algorithmic framework for performing collaborative filtering, *Proc. International on ACM SIGIR Research and Development in Information Retrieval*, (SIGIR98) 1998

[3] P. Resnick and H.R. Varian, Recommender Systems, Special Issue of *Communications of the ACM*, 40(3), 1997