COMP621U Project Proposal

Lu Wang, Hua Liu and Hao Hu

March 22, 2011

1 Introduction

People have remarkably diverse tastes in music, which reflect diversity in personalities, cultures and age groups. Recently Yahoo! Music offers a wealth of information and services related to many aspects of music, such as user ratings, which can be utilized to analyze the encoded information on how songs are grouped, which artists complement each other, and which songs users would like to listen to. Such an exciting analysis introduces new scientific challenges and also stimulates KDD Cup contest to releases two tracks:

- To predict scores that users gave to various items
- To separate loved songs from other songs

To accomplish these two tasks, we will use a collaborative filtering approach which could make automatic predictions about the interests of a user by collecting taste information from many users.

2 Datasets

Some noteworthy features of this year's KDD Cup dataset include:

- It is of a larger scale compared to other datasets: about 300M ratings
- It has a very large set of items: over 600K items
- There are four different categories of items: tracks, albums, artists and genres. These items are all linked together within a defined hierarchy.
- Timestamps for each rating is given so that session analysis of user activities can be employed.

3 Plans and Methodologies

3.1 Task 1

Task 1 is a traditional collaborative filtering task which can be viewed as link prediction or rating prediction. Some specific properties of this task include: time and session information, which can be used to analyze the temporal drift of users' taste in the rating matrix; taxonomy of items, which can be used to pass preferences of rating relationship from items that are rated in the training set to items that are unrated; genre, which serves as tags to all the items mentioned in this ontology, which can be also used to infer hidden preferences. Compared to traditional CF algorithms like the Matrix Factorization model, it is more convenient to incorporate these taxonomy information via the neighborhood algorithm perspective. Therefore, to accomplish Task 1, we will focus more on neighborhood-based algorithms instead of the MF methods.

3.2 Task 2

Task 2 is a top-K item recommendation task. We'll attempt to build our model based on several previous algorithms, including but not limited to, popularity-based and item-base models, SVD models and bipartite graph models. Such baselines are then constructed to form an ensemble for better results. The way we construct the ensemble may include Bagging or Adaboost or weighted averaging of the base learner's prediction.

3.3 Timeline

We will first focus on Task 1 in our project. If time permits, we will continue to work on Task 2. We plan to target our project in the following steps:

- Step 1 - Analyze some basic properties and statistics of the dataset, read previous works on CF problems utilizing taxonomy information and top-K item recommendation.
- Step 2 - Build the neighborhood model incorporating taxonomy information.
- Step 3 - Attempt to add taxonomy information into MF models with temporal dynamics.
- Step 4 - Implement some baselines for top-K item recommendation tasks and then build ensemble algorithms.
- Step 5 - Finalizing and submission of the project report.

4 Related Work

Much work has been published on CF in recent years [6], however, little research work has been done to incorporate taxonomy information with recommender systems. [8] is one of the earliest work on this topic. In [8], relationships between super-concepts and sub-concepts are studied and user profiles are generated based on these classifications of products that the customers have chosen. Such profiles propagate and constitute the basis of rating prediction. As an improvement of this work, [4] takes both item taxonomy and folksonomy into consideration when predicting ratings. Another important feature of Task 1 is the evolution of users' interest toward music with time. The effect of temporal dynamics has also been studied extensively in collaborative filtering. For example, in [2], the authors propose an algorithm to compute the time weights for different items in a manner that will assign a decreasing weight to old data. Besides, in [3], Yehuda provides new algorithms which could beat the previous CF algorithms on the Netflix dataset. Last but not the least, the large number of items in Task 1 makes it possible that some items are completely new in the test set. From such a perspective, some novel items must be recommended to the user even if such items have never been rated. In [5], the authors propose an approach to identify novel items that may suit the user and make recommendations. Experiments are also conducted on music recommendation tasks.

For Task 2, it is different from Task 1 since we only need to know the "best bet" recommendation and not that the predicted rating values. Such a task is usually referred to as the top-N recommendation task, where the goal of the recommender system is only to find a few specific items which are most appealing to the user. [1] provides some empirical results supporting the idea that CF algorithms that aim to optimize RMSE values are inappropriate for making top-N recommendation tasks. It can also be viewed as a collaborative ranking problem. In [7], a maximum margin matrix factorization algorithm is proposed to solve this problem.

References

[1] Paolo Cremonesi, Yehuda Koren, and Roberto Turrin. Performance of recommender algorithms on top-n recommendation tasks. In *RecSys*, pages 39–46, 2010.

[2] Yi Ding and Xue Li. Time weight collaborative filtering. In *CIKM*, pages 485–492, 2005.

[3] Yehuda Koren. Collaborative filtering with temporal dynamics. In *KDD*, pages 447–456, 2009.

[4] Huizhi Liang, Yue Xu, Yuefeng Li, and Richi Nayak. Personalized recommender system based on item taxonomy and folksonomy. In *CIKM*, pages 1641–1644, 2010.

[5] Makoto Nakatsuji, Yasuhiro Fujiwara, Akimichi Tanaka, Toshio Uchiyama, Ko Fujimura, and Toru Ishida. Classical music for rock fans?: novel recommendations for expanding user interests. In *CIKM*, pages 949–958, 2010.

[6] Xiaoyuan Su and Taghi M. Khoshgoftaar. A survey of collaborative filtering techniques. *Adv. Artificial Intellegence*, 2009, 2009.

[7] Markus Weimer, Alexandros Karatzoglou, Quoc V. Le, and Alex J. Smola. Cofi rank - maximum margin matrix factorization for collaborative ranking. In *NIPS*, 2007.

[8] Cai-Nicolas Ziegler, Georg Lausen, and Lars Schmidt-Thieme. Taxonomy-driven computation of product recommendations. In *CIKM*, pages 406–415, 2004.