Capstone Project Proposal for Team #14

Himangshu Ranjan Borah (hborah)
Rahul Prashant Shah (rshah5)
Krunal Dhanji Gala (kgala)
Siddhant Abhaykumar Doshi (sadoshi)
Sushma Ravichandran (sravich)
Harsha Kunapareddy(skunapa)

Specific description of the learning objectives:

The recommendations systems are one of the most inevitable part of business intelligence in today's world. These systems typically follow the generation of suggestions based on the explicit feedbacks from the user and implicit feedbacks associated with the user data. In the naive recommendation systems like the ones covered in class, the associations were typically extracted from the whole database of users, products and corresponding reviews. However, in the practical scenario, digging the whole network of users for generating recommendations barely makes much sense. The main reasons being the following:

- The size of the full user network is normally exponentially huge in practical business scenarios and often it's not possible to generate the association matrices on the full data as the results are expected in a blistering fast speed in the real time applications.
- Associating all the users and products for generating recommendations are often not much meaningful.

Therefore, to overcome these limitations, we frequently consider clustering the users or the products according to certain criteria and try making recommendations only for those clusters. This is a very interesting field of active research and we are planning to dig a little into the problem domain in our Capstone project. The project will cover the understanding of some of the very popular latest real time recommendation systems like the ones proposed in the papers[1][2][3][4]. We will get a clear view of how the currently being implemented large scale commercial systems like Amazon are working for such enormous amounts of data and in real time. The project will involve usage of many of the topics covered in the course like community clustering, optimizations and evaluation matrices and getting hold of the state-of-the-art libraries for implementing those.

BI Use Case:

Most Recommendation Systems co-relate high efficiency with high accuracy. On further research of this concept we realised that accuracy alone may not contribute to customer retention in a business environment. We found that the importance of notions such as

scalability and diversity to play an important role while determining the quality of the recommendation. For example, the dataset looks as follows:

User1: Batman, Superman
User2: Batman, Superman, Whispers
User3: Superman, Turbo

Suppose we have to predict a movie for User4 who we know likes Batman, then instead of recommending Superman (which we are certain that he/she will like), Turbo will be a good recommendation since User3 likes it.

In a universe of N users, if the value of N increases exponentially then it is quite taxing to apply the recommendation algorithm for the whole dataset before suggesting recommendations. At this juncture, we note that the recommendations for two users with dissimilar demographics may not affect each other. Hence, we propose a heuristic to find similar-user-clusters so that only a fraction of the total number of users is compared to a user before recommendations are generated. When it comes to the bookcrossing dataset, the intention is to suggest new books to a user from other user clusters in order to widen the scope of book suggestion and sustain the business. The same goes for the Movie Lens Dataset as well.

Datasets:

Book Crossings: Collected by Cai-Nicolas Ziegler in a 4-week crawl (August / September 2004) from the Book-Crossing community with kind permission from Ron Hornbaker, CTO of Humankind Systems. Contains 278,858 users (anonymized but with demographic information) providing 1,149,780 ratings (explicit / implicit) about 271,379 books.

MovieLens: GroupLens Research has collected and made available rating data sets from the MovieLens web site. The website has the data available in different granularities and it's considered as one of the best datasets for recommendation systems' research.

Demo:

We plan to use Python / R to implement the clustering algorithms to recommend the books. We plan to research and implement the algorithms and their ensembles from the below cited papers. Given the demographics of a particular user in Book Crossings Dataset, we plan to recommend him the books that he might be interested in based on similar users around his location and belonging to a particular age group. The similar experiments will be performed

on the MovieLens Dataset as well and performances will be evaluated from statistical standpoint.

Methods/algorithms:

We are implementing 2 algorithms:

A. First we will cluster the item and user using **ClusDiv** algorithm[1]. The distance metric we are going to use is Pearson's correlation.
B. Find the ratings of the products for the customers using the formula given in [2].
C. The evaluation metric for the above algorithm is **MAE** (Mean Absolute Error)
D. The diversity is evaluated using z-diversity as suggested in the paper[1].

Contributions:

- Harsha and Sushma will perform the clustering of the users based on the demographics.
- Himangshu and Rahul will perform the recommendation algorithm on the clusters given.
- Krunal and Siddhant will work on evaluating the model created.

References:

1) Sarwar, Badrul et al. "**Item-based collaborative filtering recommendation algorithms**." *Proceedings of the 10th international conference on World Wide Web* 1 Apr. 2001: 285-295.
2) Aytekin, Tevfik, and Mahmut Özge Karakaya. "**Clustering-based diversity improvement in top-N recommendation.**" *Journal of Intelligent Information Systems* 42.1 (2014): 1-18.
3) Pham, Manh Cuong et al. "**A Clustering Approach for Collaborative Filtering Recommendation Using Social Network Analysis.**" *J. UCS* 17.4 (2011): 583-604.
4) Braak, Paul te, Noraswaliza Abdullah, and Yue Xu. "**Improving the performance of collaborative filtering recommender systems through user profile clustering.**" *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology-Volume 03* 15 Sep. 2009: 147-150.