Data Mining Techniques for Recommending Stores to Shopping Malls

John Wu
Department of Mathematics
New York University
Courant Institute
Email: jsw452@nyu.edu

Ying Zhang
Department of Computer Science
New York University
Courant Institute
Email: zy674@nyu.edu

Purnima Padmanabhan
Department of Computer Science
New York University
Courant Institute
Email: pp1492@nyu.edu

***Abstract*—In this paper, we try to recommend stores for malls. We do this by using using various traditional recommendation techniques as well as One-Class data techniques. We compare each of these techniques and combine them together to make an ensemble of recommendation techniques.**

I. Introduction

It is of great interest for shopping mall owners to determine what store should be selected to fit an empty slot in a (target) shopping mall so that the shopping mall realizes maximum profit from this decision. We have a dataset of all the malls in the USA along with the county information of the malls [1], and all the stores of various categories within each mall. Each store falls into three groups: category (restaurant, appeals-to-parents, clothing, etc), average rating of all users and how high-ended each store in a category is (Armani, Louis Vuitton are high-ended stores in apparels, we can consider fast-food joints as low-ended stores and fine-dining is high-ended and so on). For example, a mall in a fancy neighborhood in Manhattan might want to fill an empty slot with a Michael Kors outlet, while a mall in a modest neighborhood in New Mexico might make more profit by renting out the space to a Tex-Mex food chain instead. Intuitively, the store that needs to fill in the empty slot will depend on various factors.

We used a series of recommendation system techniques to account these models and ultimately use these techniques to predict whether a store is suitable for a mall. The malls and stores can be seen as users and items in the context of the current recommendation system literature. We will apply One-Class User-based Collaborative Filtering (OCUCF) [2], One-Class Item-based Collaborative Filtering (OCICF)[2], Content-Based Recommender Systems [3] along with Weighted Alternating Least Squares (WALS) [4] with various features to recommend stores with malls. Inherently, these models have their own strengths and weaknesses [3]. We combined these models together using Ordinary Least Squares as a ensemble technique.

We will also discuss the performance of each these models using Root Mean-Squared Error (RMSE)and Mean Average Precision (MAP) to compare all of these models.

II. Motivation

As previously stated, it is important for mall owners to decide which stores will be appropriate for their malls. Intuitively, it would be best to recommend stores that are similar by some metric of current stores in a mall. Also, it is useful to recommend complements of stores in order to attract more customers. Techniques such as user-based collaborative filtering can utilize store and mall information to find the inter-similarity of stores and malls. This inter-similarity can be used to appropriate stores to malls. Also, techniques such content-based recommendation systems can find the similarity between stores and malls to recommend stores that are similar to a mall itself.

Additionally, these recommendation systems are not only useful for shopping mall managers, but it is also recommend stores for individuals as well. Individuals can be represented as malls with their favorite stores as stores contained in an arbitrary mall and their geographical location as the malls location. Using the proposed ensemble recommendation system can inform users of possible stores they have not seen before and can be used in popular websites, such as Yelp.

III. Related Work

This work is an extension of the work done by [6]. [6] used the pairwise distance between malls as comparison features for creating a user-based collaborative filter to predict the number of stores of a certain category. We extend upon this by actually predicting an appropriate store for a mall.

For our dataset, we are not dealing with this issue since our dataset is a One-Class dataset, a dataset where all entries are binary and all every entries are filled in [4]. Negative examples and missing positive examples are embedded as 0s. Our goal is to change appropriate 0s as 1s using techniques such as WLAS [4] and traditional recommendation system techniques in the context of One-Class problems. [2]

[3] discusses the underlying basis of recommendation systems. Particularly, [3] discusses how traditional recommendation systems impute the values of missing data. These techniques do not directly apply to our dataset as well will discuss further about this in the next section.

Another popular technique that can be used to recommend stores to malls are supervised learning for regression. In the context of real-estate, [7] uses features that have been selected to measure the housing value of the Boston suburb by using techniques such as Support Vector Machines (SVM), Partial Least Squares (PLS), and Least Squares Support Vector Machine (LSSVM). Seeing the performance of these techniques,

we implemented Logistic Regression as a baseline method to compare our ensemble technique.

Using appropriate features for our recommendation models are important. [8] uses median-price indices, repeat-sales indices, hedonic indices, and stock-market-based indices to predict real estate prices. Similarly, for this project, we used demographic data [9], which consists of the racially composition of each county, as well as the industry data [1] of each county. These county information are used as features for the collaborative filtering model to represent the user features of the malls.

IV. Background and Methodology

1) Collaborative Filtering: Collaborative Filtering has been discussed multiple times in the motivation related work section of this paper. To formalize the idea of user-based Collaborative Filtering, it is a system that makes new predictions by first finding users with similar ratings to other user and then computes the weighted average of their ratings. This is essentially done with an arbitrary metric that relates the features of user x and user y. More formally, let $U = (u_{jf})$ be the matrix that refers to the ever user's features where (u_{jf}) represents the f feature of user j, $S = (s_{ij})$: be the similarity matrix between user i and j, and $X = (x_{ij})$: be the item-user Matrix. Notice that if U is given, S is computed by a function f, where $s_{ij} = f(U_{i.}, U_{j.})$.

$$X_{ij} = \overline{X_{.j}} + \frac{\Sigma_{l \in N(j;i)} s_{il}(X_{il} - \overline{X_{.j}})}{\Sigma_{l \in N(j;i)} s_{il}} \tag{1}$$

Note that $\overline{X_{.j}}$ is the average rating user j has made to the items that they rated and $N(j;i)$ is the set of users that rated i and is in the same neighborhood as user j. We denote neighborhood as the top k most similar users with respect to j.

Furthermore, an item-based system predictions are made by first finding items with similar ratings to other items and then computes the weighted average of their ratings. This is defined as:

$$X_{ij} = \overline{X_{i.}} + \frac{\Sigma_{l \in N(i;j)} s_{li}(X_{lj} - \overline{X_{i.}})}{\Sigma_{l \in N(i;j)} s_{li}} \tag{2}$$

Note that $\overline{X_{i.}}$ is the average rating for item i for all the users that rated i and that $N(i;j)$ is the set of items that are that are related to i for the items that j rated. The neighborhood set scheme discussed about user-based collaborative filtering applied for items.

A. Matrix Factorization and Weighted Alternating Least Squares

As discussed previously, Collaborative Filtering has been commonly used for recommendation problems. Matrix Factorization Methods are used [**?**] to address the other type of recommendation problem, One-Class Data. Factorization techniques are sometimes more favorable than Collaborative Filtering techniques because they can reduce the RMSE than Collaborative Filtering. More formally, let X be the original rating matrix and U and V be the low rank approximations of the matrix. The problem is defined as [**?**].

$$X \approx UV \tag{3}$$

According to [**?**], it is important to include weights to compute the approximation of X for OCCF. Thus, the problem can be defined as:

$$\underset{R}{\text{minimize}} \quad \underset{ij}{\Sigma} W_{ij}(R_{ij} - Xij)^2 \tag{4}$$

The reason why we include weights is that we are confident that some of the values X_{ij} must hold true. We know that a user likes a item since X_{ij} is 1, so we must give it high penalization weight W_{ij} if it is the prediction $(UV)_{ij}$ is greatly different from X_{ij}.

$$\underset{U,V}{\text{minimize}} \quad \underset{ij}{\Sigma} W_{ij}((UV)_{ij} - Xij)^2 + \lambda(\|U_{i.}\|_2^2 + \|V_{.j}\|_2^2)$$

A more throughout discussion of how to solve this problem can be seen in [**?**].

V. Design and Implementation

A. Data Processing

After the data collection, lots of data cannot be used directly for analyze, we need to filter the data and join them together for better analyze. We use Apache Pig and Python to process data.

Firstly, we process the stores list. In our store list, There are three type of problems, the first one is Lots of Stores are not store, for example: atm, vending machines , community service etc. we need to delete these items. Second one is Lots of same Stores have same names, for example: Starbucks and Starbucks Coffee, we need to keep the name the same . Last one is Lots of stores have special characters and misspelling, we need to correct these.In order to get the best predict result, we create three different regular expression rules to filter our data and create unique id for every store. We Write Pig Script to load data first, then use a pig use-define function to filter these data. Secondly, after we get the clean store list, we use a yelp crawler to get all there store data. Data Format: store name ,category , level of expense,popularity rating. Thirdly, Every County has different industry category, (the category is influenced by the owner, industry and time), some type of industry is very rare, just existing n some special places. These items will influence the accuracy of the prediction.To solve this problem, we Find the intersection of all counties industry, and just keep the common industry in the data.After filter all mall data, we need to join all geographic data , industry data and mall data together, so that in machine earning part ,we can create mall fearture matrix from these data.Join Store list with yelp data as the store data. Now we have filter our data and join them together, we can use these data to analyze

	Pos Examples	Neg Examples
Uniform (Rong 2008)	1	$W_ij = \delta$
Mall-Oriented (Rong 2008)	1	$W_ij \propto \Sigma_j X_{ij}$
Store-Oriented (Rong 2008)	1	$W_ij \propto 1 - \Sigma_i X_{ij}$

TABLE I. FINDING WEIGHTS

B. Methods

1) One-Class Collaborative Filtering: Even though traditional collaborative filtering methods are do not directly solve the one-class problem, we modified it to solve the problem. [2] has previously done this and relatively good results for doing this. We modified User-based Collaborative Filtering by doing the following:

- Compute $\forall s_{ij} \in S$: $s_{ij} = f(U_{i.}, U_{j.})$
- $\forall\ X_{ij} = 0$
  - Compute $\hat{X_{ij}} = \frac{\Sigma_{l \in N(j;i)} s_{il} X_{il}}{\Sigma_{l \in N(j;i)} s_{il}}$

  $X_{ij} = \hat{X_{ij}}$, $\forall\ X_{ij} = 0$

As mentioned previously, the features U that have been used are the demographic and distance information.

The item-based Collaborative Filter can be derived similarly. The features that were used are the yelp data.

2) One-Class Matrix Factorization: For the One-Class WLAS, appropriate weights need to be chosen. We experimented with the weights describe in Table 1 for our dataset, which is discussed in [4].

Notice, to experiment with these values, we only needed the binary Mall-Store binary Matrix.

VI. RESULT

VII. ACKNOWLEDGMENT

We would like to thank Roy Lowrance and Dennis Shasha for giving us advice for this project as well as the opportunity to continue upon the work of this mall dataset, which was done by Joe Jean who is responsible for information of malls and the stores that are contained in the malls. We would also like to thank Suzanne McIntosh for giving us the opportunity to use Amazon Web Services for running our dataset and algorithms on a cluster. Also, we would like to thank her for giving us advice on how to implement different big data technologies on these clusters.

VIII. FUTURE WORK

Subsection text here.

REFERENCES

[1] B. of Labor and Statistics, "Quarterly census of employment and wages," 2013.

[2] Y. Li, J. Hu, C. Zhai, and Y. Chen, "Improving one-class collaborative filtering by incorporating rich user information," in *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM, 2010, pp. 959–968.

[3] A. Rajaraman and J. D. Ullman, *Mining of massive datasets*. Cambridge University Press, 2011.

[4] R. Pan, Y. Zhou, B. Cao, N. N. Liu, R. Lukose, M. Scholz, and Q. Yang, "One-class collaborative filtering," in *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*. IEEE, 2008, pp. 502–511.

[5] S. Lai, Y. Liu, H. Gu, L. Xu, K. Liu, S. Xiang, J. Zhao, R. Diao, L. Xiang, H. Li *et al.*, "Hybrid recommendation models for binary user preference prediction problem." in *KDD Cup*, 2012, pp. 137–151.

[6] J. Jean, R. Lowrance, and D. Shasha, "Store-mall recommendation," http://ml2014.herokuapp.com/, accessed: 2015-02-01.

[7] J. Mu, F. Wu, and A. Zhang, "Housing value forecasting based on machine learning methods," in *Abstract and Applied Analysis*, vol. 2014, 2014.

[8] E. Ghysels, A. Plazzi, W. N. Torous, and R. I. Valkanov, "Forecasting real estate prices," *Handbook of economic forecasting*, vol. 2, 2012.

[9] U. Concensus, "Demographic informaion," 2013.

[10] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2001, pp. 556–562.

[11] Y. Zhang, J. Wu, and P. Padmanabhan, "Mall recommendation presentation," github.com/lily-zhangying/find_best_mall/blob/master/diagrams.pdf, accessed: 2015-04-02.