A Brick-and-mortar Store Recommendation System based on Online Shopping Behavior

Ye Wang, Shihao Wang
Northeastern University
Instructor: Tina Eliassi-Rad

Introduction

- What problem are you solving?

  As mobile devices become ubiquitous in our daily life, people are getting more comfortable sharing their real-time locations with various location-based services, such as navigation, car ride hailing, restaurant/hotel booking, etc. On the other hand, huge amount of user data has been accumulated in online services like online shopping. We may explore the value of users' behavior to help the public find their demand in location based services using data mining techniques.

  Concretely, we want to recommend the potential merchants that users will visit in the future, based on the user behavior records of online website and on-site merchants.

- What is your problem definition?

  For this recommendation task, the input consists of two parts. First, the behavior records(click, buy, etc.) of a user from online websites. Second, the user's location represents where is the user now. And we output an ordered merchant list including the nearby stores the user may like.

- Why is it interesting and important?

  Generally speaking, a large online retail platform serves far more than people than on-site merchants, the online platform will accumulate lots of valuable user data and utilize it for improving user experience, such as personalized recommendation and search. However, it is difficult for brick-and-mortar stores. First, the number of people visiting a store is limited compared to an online platform. Second, it's relatively hard for them to collect user behavior data. Therefore, it is very meaningful to explore users' online shopping behavior for providing them more accurate location-based store recommendation, especially when users enter new areas they rarely visited in the past.

- Why is your problem hard?

  First, the key point of our work is to mine latent correlations between merchants and online sellers. We cannot directly measure the distance between merchants and online sellers because they don't lie in the same feature space. Such correlations are hard to define because they are under control by lots of latent variables. Second, Extracting effective features for measuring the distance or similarity of different users is also important but difficult. Because the meta data doesn't give us many useful features directly. We need to construct them by analyzing multiple tables jointly. Third, recommendation strategy also affects the final results highly and should be designed carefully.

Data

- What data set are you going to use and why is it appropriate?

  The involved training data are shown as follows:

  1. Taobao Online user behavior dataset, which provides users and 44528127 purchase records.(Table 1)

  2. Koubei Brick-and-mortar users shopping records dataset, which provides 230496 users and 1081724 purchase records, and 5910 on-site merchants.(Table 2 & 3)

  Taobao.com is the largest online retail platforms in China, serves for more than 10 million merchants and over 300 million customers. Meanwhile, Koubei offers restaurant and retail store recommendation and payment services for a huge number of customers. A user enjoying services provided by these two groups often has a unified online account. So these datasets are adequate and representative. We are excited for mining behavior and interests of such a large number of users.

Field	Description
User_id	unique user id
Seller_id	unique online seller id
Item_id	unique item id
Category_id	unique category id
Online_Action_id	"0" denotes "click" while "1" for buy
Time_Stamp	date of the format "yyyymmdd"

Table 1: Online user behavior on taobao.com

Algorithms

- What are the key components of your approach?

  Our approach has five basic modules: 1. Feature engineering. We will conduct feature engineering for users,

Field	Description
Merchant_id	unique merchant id
Location_id	unique location id
Time_Stamp	date of the format "yyyymmdd"

Table 2: Users shopping records at brick-and-mortar stores

Field	Description
Merchant_id	unique merchant id
Location_id_list	available location list, e.g. 1:356:89

Table 3: Merchant information

merchants and sellers separately (for convenience, we denote online stores as sellers and brick-and-mortar stores as merchants). We will extract several features that are useful for measuring the similarity of users/merchants/sellers. 2. Similarity matrix transformation. We will apply the kernel method to transfer feature matrices into kernel matrices, where $k_{user}(i, j)$ can be regarded as the similarity between user i and j.

3. Spectral Clustering [Ng, Jordan, and Weiss2002]. Spectral Clustering gives us the cluster results and the clustering centroids' distribution of users, merchants and sellers separately.

4. Correlation computing between Merchants and Sellers. We plan to take advantage of users as the intermediary. We can use soft alignment to compute the correlation between one user cluster and one merchant cluster as follows:

$$R(C_u, C_m) = \frac{\sum_{m \in M} P(m|C_m) \frac{\sum_{u \in U} P(u|C_u) R(u,m)}{\sum_{u \in U} P(u|C_u)}}{\sum_{m \in M} P(m|C_m)} \quad (1)$$

, where U and M are the set of user and merchant clusters. $R(u, m)$ is the purchase frequency of user u to merchant m. The same computational process can be applied to the correlation between user and seller clusters. Finally, we can generate the similarity score between one merchant cluster and one seller cluster as follows:

$$R(C_s, C_m) = \frac{\sum_{u \in U} R(C_u, C_m) R(C_u, C_s)}{Z} \quad (2)$$

, where Z is a normalization factor.

5. Collaborate Filtering for recommendation. The final recommendation is straightforward. For the user who clicks or purchases certain online sellers, we find the merchants with top similarity scores from given merchants set to form the ranking list.

- Why are these algorithms appropriate?

  We adopt kernel method to build up the similarity matrix because the kernel matrix can be regarded as the similarity matrix. Moreover,if the kernel is properly chosen, we can obtain a good measure of similarity in the kernel space. We use Spectral Clustering mainly because it can capture more complex cluster shapes and have a good performance on sparse data.

Evaluation

- What metrics will be used to validate the approach?

  We plan to conduct a three-steps evaluation for our work. The first stage is to evaluate the clustering performance. The following internal metrics can be considered:

  – Validity via Correlation (Proximity and Incidence Matrices)
  – Cohesion and Separation, Silhouette coefficient
  – Comparison with other clustering algorithms

  The second stage would be evaluating the confidence of our merchant-seller similarity matrix. As our most important output, the confidence of this matrix will directly determine the performance of our recommendation. We plan to use some visualization method to see if there is any consistence between the spatial distance and similarity. We are still thinking about the better ideas.

  The third stage is to evaluate the recommendation result. We can definitely hold out some labeled data as our test dataset. However, we still expect to have some unsupervised metrics to tackle with it.

References

[Ng, Jordan, and Weiss2002] Ng, A. Y.; Jordan, M. I.; and Weiss, Y. 2002. On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*, 849–856.