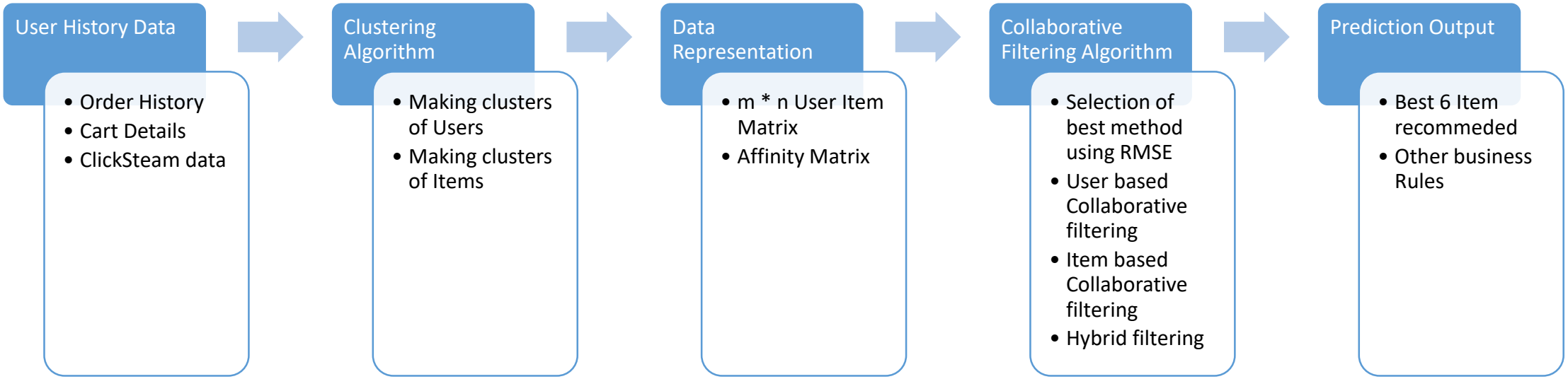
Sample Recommendation Systems

Collaborative filtering Approaches

Darshil Gohel
Masters Student
University of Texas at Dallas

Technical Architecture 1



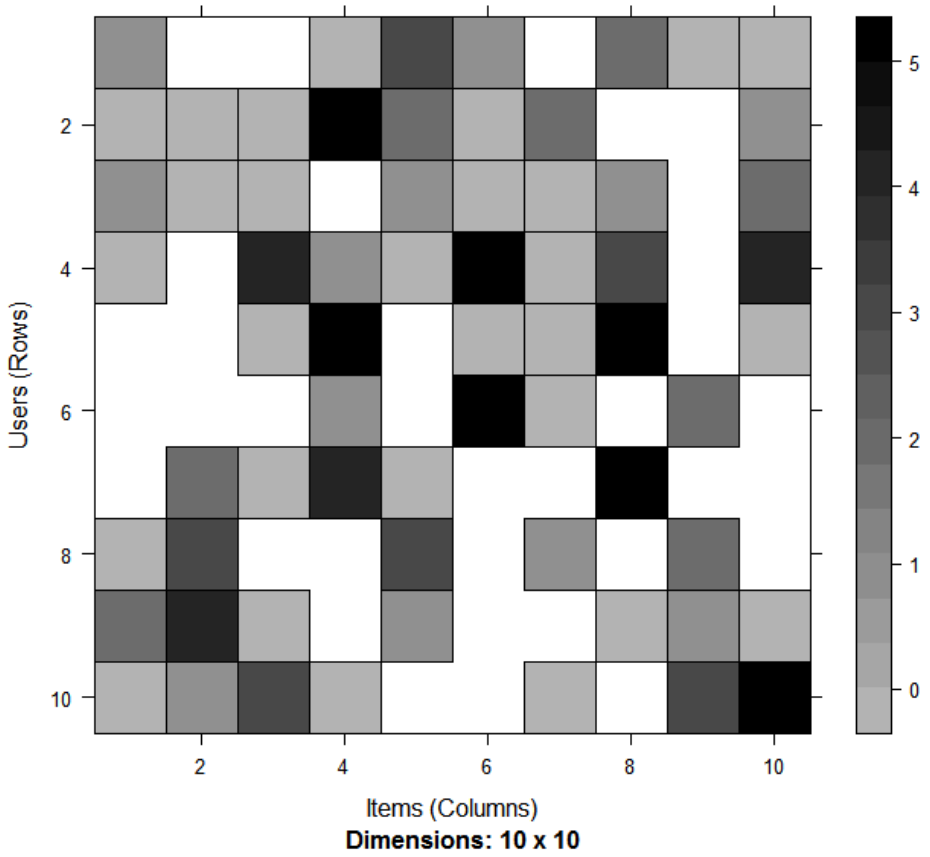
- Recommendation system in real time are far more complex
- Complexity has been greatly reduced to make this sample recommendation system

Technical Architecture 2

- **User History Data**: Order History data, cart details , clickstream data of the user
- **Clustering Algorithm**: This algorithm(e.g K means clustering algorithm) divides the user and items into certain categories. Like Men or Women, Age groups, color bias. Etc.
- **Data Representation**: Data has to be organized in Affinity Matrix or m*n User Item Matrix.
- **Collaborative Filtering Algorithm**: following algorithms has been selected as filtering algorithms.
  - User Based Collaborative Filtering Algorithm
  - Item Based Collaborative Filtering Algorithm
  - Hybrid Filtering Algorithm
- **Prediction Output:** Using among best algorithm chosen from above three and according to other business rules , best 6 items which is predicted for particular user.

Affinity Matrix

	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10
U1	1	NA	NA	0	3	1	NA	2	0	0
U2	0	0	0	5	2	0	2	NA	NA	1
U3	1	0	0	NA	1	0	0	1	NA	2
U4	0	NA	4	1	0	5	0	3	NA	4
U5	NA	NA	0	5	NA	0	0	5	NA	0
U6	NA	NA	NA	1	NA	5	0	NA	2	NA
U7	NA	2	0	4	0	NA	NA	5	NA	NA
U8	0	3	NA	NA	3	NA	1	NA	2	NA
U9	2	4	0	NA	1	NA	NA	0	1	0
U10	0	1	3	0	NA	NA	0	NA	3	5



Where U1,U2→ Users

I1,I2→Items

3,2,4→ rating given by Users to some items.

Selection of Collaborative Filtering Method -An Overview

Following type of Users have been considered:

- New
- Returning
- Loyal

Following methods have been compared:

- Item Based Collaborative Filtering
- User Based Collaborative Filtering
- Hybrid (50% Item Based Collaborative Filtering + 50% User Based Collaborative Filtering)

User based Collaborative Filtering

- Basic Idea – Find other users whose past rating behavior is similar to that of the current user and use their rating on other items to predict what the current user will like.
- Required: Ratings Matrix and similarity function that computes the similarity between two users.
- The selection of neighbors can be random or based on the thresold value.
- User U's prediction for item I is given by Pu,I

$$p_{u,i} = \bar{r}_u + \frac{\sum_{u' \in N} s(u,u')(r_{u',i} - \bar{r}_{u'})}{\sum_{u' \in N} |s(u,u')|}$$

Item Bases Collaborative Filtering

- Basic Idea: Recommend items that are similar to user's highly preferred items.
- Provides performance gains by lending ifself well to precomputing similarity matrix.
- User U's prediction for item is given by Pu,I

$$p_{u,i} = \frac{\sum_{j \in S} s(i,j) r_{u,j}}{\sum_{j \in S} |s(i,j)|}$$

- Cosine similarity or conditional probability is used to comput item item similarity.

Hybrid Collaborative Filtering

- It is combination of user bases and item based collaborative filtering methods in different proportion
- Sometimes , Hybrid collaborative filtering is more recommended to get best results
- General proportion of tow methods :
  - User based collaborative filtering (50%)
  - Item based collaborative filtering (50%)

Selection of Collaborative Filtering Method -Assumptions

- For comparison of methods, Affinity Matrix have been created according to different case of new, returning and loyal customers.
- Rating Scale → 0 to 5 (Total 6 values)
- The RMSE (Root Mean Square error) is calculated for all the methods and selected the best method according to lowest RMSE.
- Datasets (affinity matrices) have been divided as 90% of training data and 10% as testing data for comparison of different techniques.

Selection of Collaborative Filtering Method - For New Users

- Assumptions:
  - Number of Users: It is assumed here that the user is completely new and system don’t know about that except the clickstream data. Hence , clustering system which the prior step of the recommendation system will failed to categorized it and hence New User has to be compared with large number of users. So, we are taking number of user as large as possible.

  - Number of Items: If I would be new user to zappos site then I will navigate the zappos site with curiosity of knowing which items have been sold rather then just jumping of particular category. So , in case of new User, they will try to explore more and more items and item categories then the loyal customer. New users are confused or still in dilemma what to buy and what to not. Hence keeping more and more item in our considering and hence our itemset for new user will be as large as possible.

Selection of Collaborative Filtering Method - For New Users

- Assumptions:
  - Rating of Items: from clickstream data, for new users, if user spend more time on particular item, it is more likely that he likes that product and will try to buy that product in the future.

Time spent on item(Seconds)	Rating
<10	0
10 to 20	1
20 to 30	2
30 to 40	3
40 to 50	4
>50	5

  - For existing and returning users, rating is actually rating which customer have given to particular item. We are actually comparing all the existing and new users in our cases based upon this criteria.

Selection of Collaborative Filtering Method - For New Users

- Affinity Matrix Creation Function in R:

```
m1<- matrix(sample(c(NA,0:5),1000000, replace=TRUE,
prob=c(0.50,0.30,0.15,0.02,0.01,0.01,0.01)),
        nrow=1000, ncol=1000, dimnames = list(
          user=paste('U', 1:1000, sep=''),
          item=paste('I', 1:1000, sep='')
        ))
```

Selection of Collaborative Filtering Method - For New Users

- Affinity Matrix Parameters:
  - Number of Users: As large as possible, hence nrow=1000
  - Number of Items: As large as possible, hence ncol=1000
  - Probabilities of Ratings

Rating	Probabilities	Reason
NA	0.5	it is assumed that we don't have any data for user item rating for around 50% of values
0	0.3	It is assumed that New user is just exploring site and his tendency would be exploring most of items in shot time
1	0.15	It is assumed that New user is just exploring site and his tendency would be exploring most of items in shot time
2	0.02	it is very less likely that user will stuck to particular item and spend time
3	0.01	It is almost impossible to like any particular item on first visit on site
4	0.01	It is almost impossible to like any particular item on first visit on site
5	0.01	It is almost impossible to like any particular item on first visit on site

Selection of Collaborative Filtering Method - For New Users

- Method:
  - Generated Affinity Matrix is our represented data which is feed to filtering systems
  - Entire dataset has been divided in following way:
    - Training Dataset(90% of dataset)→ To train the Model
    - Testing Dataset(10% of dataset)→ To Test the remaining data
  - Following parameters have been calculated and compared
    - RMSE(Root Mean Square Error)
    - RSE(Root Square Error)
    - RAE(Root Absolute Error)

Selection of Collaborative Filtering Method - For New Users

- Results:
  - Error calculations for the three methods:

	RMSE	MSE	MAE
UBCF	1.062457	1.128815	0.7472459
IBCF	1.102898	1.216383	0.7666330
HYBRID	1.070036	1.144977	0.7503248

  - Using above matrix, it is apparent that UBCF method is best as it has least Root Mean Square Error (RMSE) comparing to other two records.

Selection of Collaborative Filtering Method - For New Users

- Best Predicted 6 items for recommendation for user 5(U5) using UBCF method

No	Item Name
1	I736
2	I256
3	I507
4	I936
5	I486
6	I275

Selection of Collaborative Filtering Method - For New Users

- Twist
  - Zappos' primary selling base is shoes, which accounts for about 80% of its business.
  - Even though, Zappos expanded their inventory in 2007 to include clothing, handbags, eyewear, watches, and kids' merchandise and still in 2017 those items currently account for 20% of annual revenues.
  - As conclusion, it is likely that customer is looking zappos as footwear site rather than other items.
  - Hence, it is most likely that the predicted 6 items will be footwear only(which is current case in Zappos Site !!!!).
  - We should also include items (mostly one or two out of six) like cloths, handbags, watches which is other then footwear.
  - Some kind of algorithm is needed at this stage to override other prediction and add items which are other then footwear.

Selection of Collaborative Filtering Method - For New Users

- Twist
  - Assuming this algorithm is created and now final list of predicted items:

No	Item Name
1	I736
2	I256
3	I507
4	I936
5	I788
6	I999

  - Where, item 1 to 4 are the same as predicted by our recommendation algorithm and item 5 an item 6 are items which are added by our newly created algorithm and which is other then footwear.

Selection of Collaborative Filtering Method - For Returning Users

- Assumptions:
  - Number of Users: This user has already purchased one or two items from the site and system have some data about it like age group, gender, affinity of some categories of items. System also have some clickstream data. Hence,our clustering system will able to categorize the user and our returning user will fall in particular category and hence we will only compare returning user with the user which are present in that category. So, number of users will be somewhat less in this case.
  - Number of Items: Our returning user have bought some items but he is still unware of other items which are being sold on Zappos. So , in case of returning User, they will try to explore more and more items and item categories then the loyal customer.We need to recommend him more and more items. Hence keeping more and more item in our considering and hence our itemset for new user will be as large as possible.

Selection of Collaborative Filtering Method - For Returning Users

- Assumptions:
  - Rating of Items: Returning user have already rated some of our items
  - Moreover, from clickstream data, for returning users, if user spend more time on particular item, it is more likely that he likes that product and will try to buy that product in the future.

Time spent on item(Seconds)	Rating
<10	0
10 to 20	1
20 to 30	2
30 to 40	3
40 to 50	4
>50	5

  - For existing and returning users, rating is actually rating which customer have given to particular item. We are actually comparing all the existing and new users in our cases based upon this criteria.

Selection of Collaborative Filtering Method - For Returning Users

- Affinity Matrix Creation Function in R:
- ***m1<- matrix(sample(c(NA,0:5),500000, replace=TRUE, prob=c(0.40,0.20,0.10,0.10,0.10,0.05,0.03)),***

  nrow=500, ncol=1000, dimnames = list(

  user=paste('U', 1:500, sep=''),

  item=paste('I', 1:1000, sep='')

  ))

Selection of Collaborative Filtering Method - For Returning Users

- Affinity Matrix Parameters:
  - Number of Users: nrow=500
  - Number of Items: As large as possible, hence ncol=1000
  - Probabilities of Ratings

Rating	Probabilities	Reason
NA	0.40	it is assumed that we don't have any data for user item rating for around 40% of values
0	0.20	Returning User know some of the item sold on zappos and he don't like them and giving them 0 rating. It is assumed that New user is just exploring site and his tendency would be exploring most of items in shot time
1	0.10	It is assumed that returning user have given 1 rating to 10% which he has visited and data collected from clickstream data
2	0.10	It is assumed that returning user have given 1 rating to 10% which he has visited and data collected from clickstream data
3	0.10	It is assumed that returning user have given 2 rating to 10% which he has visited and data collected from clickstream data
4	0.05	Still our returning user is comparatively new to our site so he will give 4 rating to very few items
5	0.03	Still our returning user is comparatively new to our site so he will give 5 rating to very few items

Selection of Collaborative Filtering Method - For Returning Users

- Method:
  - Generated Affinity Matrix is our represented data which is feed to filtering systems
  - Entire dataset has been divided in following way:
    - Training Dataset(90% of dataset)→ To train the Model
    - Testing Dataset(10% of dataset)→ To Test the remaining data
  - Following parameters have been calculated and compared
    - RMSE(Root Mean Square Error)
    - RSE(Root Square Error)
    - RAE(Root Absolute Error)

Selection of Collaborative Filtering Method - For Returning Users

- Results:
  - Error calculations for the three methods:

	RMSE	MSE	MAE
UBCF	1.571777	2.470484	1.357055
IBCF	1.603681	2.571792	1.371418
Hybrid	1.570091	2.465185	1.356057

  - Using above matrix, it is apparent that Hybrid method is best as it has least Root Mean Square Error (RMSE) comparing to other two records.

Selection of Collaborative Filtering Method - For Returning Users

- Best Predicted 6 items for recommendation for user 5(U5) using Hybrid Method

No	Item Name
1	I481
2	I609
3	I831
4	I265
5	I214
6	I433

Selection of Collaborative Filtering Method - For Returning Users

- Twist
  - Zappos' primary selling base is shoes, which accounts for about 80% of its business.
  - Even though, Zappos expanded their inventory in 2007 to include clothing, handbags, eyewear, watches, and kids' merchandise and still in 2017 those items currently account for 20% of annual revenues.
  - As conclusion, it is likely that customer is looking zappos as footwear site rather than other items.
  - Hence, it is most likely that the predicted 6 items will be footwear only(which is current case in Zappos Site !!!!).
  - We should also include items (mostly one or two out of six) like cloths, handbags, watches which is other then footwear.
  - Some kind of algorithm is needed at this stage to override other prediction and add items which are other then footwear.

Selection of Collaborative Filtering Method - For Returning Users

- Twist
  - Assuming this algorithm is created and now final list of predicted items:

No	Item Name
1	I481
2	I609
3	I831
4	I265
5	I788
6	I999

  - Where, item 1 to 4 are the same as predicted by our recommendation algorithm and item 5 an item 6 are items which are added by our newly created algorithm and which is other then footwear.

Selection of Collaborative Filtering Method - For Loyal Users

- Assumptions:
  - Number of Users: This user is loyal user. System have all the information about this user and it is most likely that our system will easily categorize this user and this user will be compared with the same category only and so taking fewer user number into account
  - Number of Items: This customer is loyal and it is assumed that loyal customer tend to buy the item from some few number of categories only, not all categories!!Hence,it is better to recommender him the item from those categories only in which he is most loyal to sustain its loyalty rather then just throwing some random items to him. So, taking into consideration only items of most liked categories→ Fewer no of items.
  - Rating of Items: Returning user have already rated some of our items .Here, we are no more predict on clickstream data. This is prediction is only based the data system knows about the Loyal customer.

Selection of Collaborative Filtering Method - For Loyal Users

- Affinity Matrix Creation Function in R:
- ***m1<- matrix(sample(c(NA,0:5),250000, replace=TRUE, prob=c(0.25,0.10,0.10,0.15,0.20,0.10,0.10)),***

```
nrow=500, ncol=500, dimnames = list(
  user=paste('U', 1:500, sep=''),
 item=paste('I', 1:500, sep='')
))
```

Selection of Collaborative Filtering Method - For Loyal Users

- Affinity Matrix Parameters:
  - Number of Users: nrow=500
  - Number of Items: ncol=500
  - Probabilities:

Rating	Probabilities	Reason
NA	0.25	it is assumed that our loyal customer knows around 75% items and he have reated them but we don't have any data for user item rating for around 25% of values
0	0.10	Our customer is loyal and it is most likely that there will be less products which he didn't like from his favorite category.
1	0.10	Our customer is loyal and it is most likely that there will be less products which he didn't like from his favorite category.
2	0.15	There will be some items to which customer will give 2 rating
3	0.20	Most of the items our customer will like will likely have 3 rating
4	0.10	There will be some items which is loved by customer and hence customer will give them 4 rating
5	0.10	There will be some items which is loved by customer and hence customer will give them 5 rating

Selection of Collaborative Filtering Method - For Loyal Users

- Method:
  - Generated Affinity Matrix is our represented data which is feed to filtering systems
  - Entire dataset has been divided in following way:
    - Training Dataset(90% of dataset)→ To train the Model
    - Testing Dataset(10% of dataset)→ To Test the remaining data
  - Following parameters have been calculated and compared
    - RMSE(Root Mean Square Error)
    - RSE(Root Square Error)
    - RAE(Root Absolute Error)

Selection of Collaborative Filtering Method - For Loyal Users

- Results:
  - Error calculations for the three methods:

	RMSE	MSE	MAE
UBCF	1.547359	2.394320	1.281912
IBCF	1.570394	2.466138	1.294071
HYBRID	1.543242	2.381595	1.281164

  - Using above matrix, it is apparent that Hybrid method is best as it has least Root Mean Square Error (RMSE) comparing to other two records.

Selection of Collaborative Filtering Method - For Loyal Users

- Best Predicted 6 items for recommendation for user 5(U5) using Hybrid Method

No	Item Name
1	I481
2	I609
3	I831
4	I265
5	I214
6	I433

Thank you