DATA ANALYTICS

Unit 4: Collaborative Filtering System

Jyothi R.
Department of Computer Science and Engineering

Domain-specific Challenges in Recommender Systems

1. Context-Based Recommender Systems: It could include time, location, or social data. For example, the types of clothes recommended by a retailer might depend both on the season and location of the customer. Even particular type of festival or holiday affects the underlying customer activity.

2. Time-Sensitive Recommender Systems:

i. The rating of an item might evolve with time, as community attitudes evolve and the interests of users change over time. User interests, likes, dislikes, and fashions inevitably evolve with time.

ii. The rating of an item might be dependent on the specific time of day, day of week, month, or season.

iii. For example, it makes little sense to recommend winter clothing during the summer, or Raincoats during the dry season.

Domain-specific Challenges in Recommender Systems

3. Location-Based Recommender Systems
    i) User-Specific Locality
    ii) Item-specific Locality

4. Social Recommender Systems
    i) Structural Recommendation of Nodes and Links
    ii) Product and Content Recommendations with social influence.
    iii)Trustworthy Recommender Systems
    iv) Leveraging Social Tagging Feedback for Recommendations

DATA ANALYTICS

Advanced Topics and Applications

1. The Cold-Start Problem in Recommender Systems: Most people have not rated most items - Cold Start:
   - New items have no ratings
   - New users have no history

2. Attack-Resistant Recommender Systems

3. Group Recommender Systems

4. Multi-Criteria Recommender Systems

5. Active Learning in Recommender Systems

6. Privacy in Recommender Systems

7. Application Domains

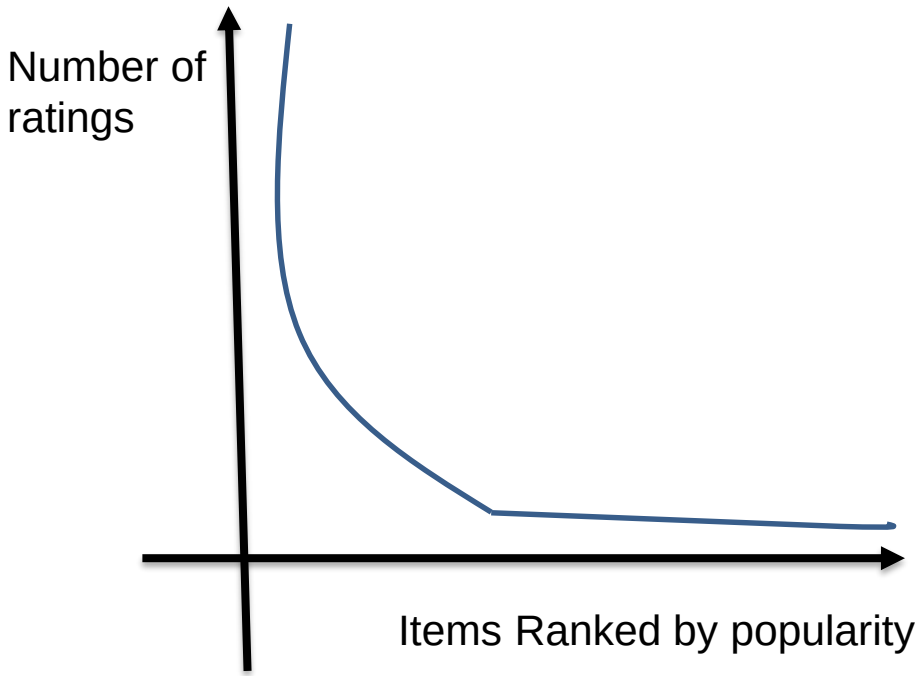
From Scarcity to Abundance

- Shelf space is a scarce commodity for traditional retailers TV networks, Movie Theatres.
- More choice necessitates better filters.
- The web enables near-zero-cost dissemination of information about products.

From scarcity to abundance gives rise to the “Long Tail” phenomenon.

DATA ANALYTICS

The long tail:



The long tailed distribution implies that the items, which are frequently rated by users, are fewer in number. This fact has important implications for neighborhood-based collaborative filtering algorithms because the neighborhoods are often defined on the basis of these frequently rated items. In many cases, the ratings of these high frequency items are not representative of the low-frequency items because of the inherent differences in the rating patterns of the two classes of items. As a result, the prediction process may yield misleading results.

The economics of abundance:

Items might be a books, music, videos, or news articles

DATA ANALYTICS

Overview of recommendations

1. Editorial and hand curated
   i) List of favourites
   ii) List of “essential” items

2. Simple aggregates.
   i) Top 10 Most Popular
   ii) Most recent uploads

3. Tailored to individual users
   i) Amazon
   ii) Netflix
   iii) Pandora's

Formal Model

Utility Function : A function that looks at every pair of a customer and item and maps it.

U: C x S → R

I. C= set of customers and
II. S=Set of items
III. R=Set of ratings, it is a totally ordered set eg.-5 stars, real no in[0..1]

Utility Matrix

	Avatar	KGF	Matrix	Bahubali
Alice	1		0.2	
Bob		0.5		0.3
Carol	0.2		1	
David				0.4

Key Problems

1. Gathering “known” ratings for matrix: How to collect the data in the utility matrix?

2. Extrapolate unknown ratings from the known ones: mainly interested in high unknown ratings. i.e., we are not interested in knowing what you don’t like but what you like

3. Evaluating extrapolation methods: how to measure success/performance of recommendation methods?

DATA ANALYTICS

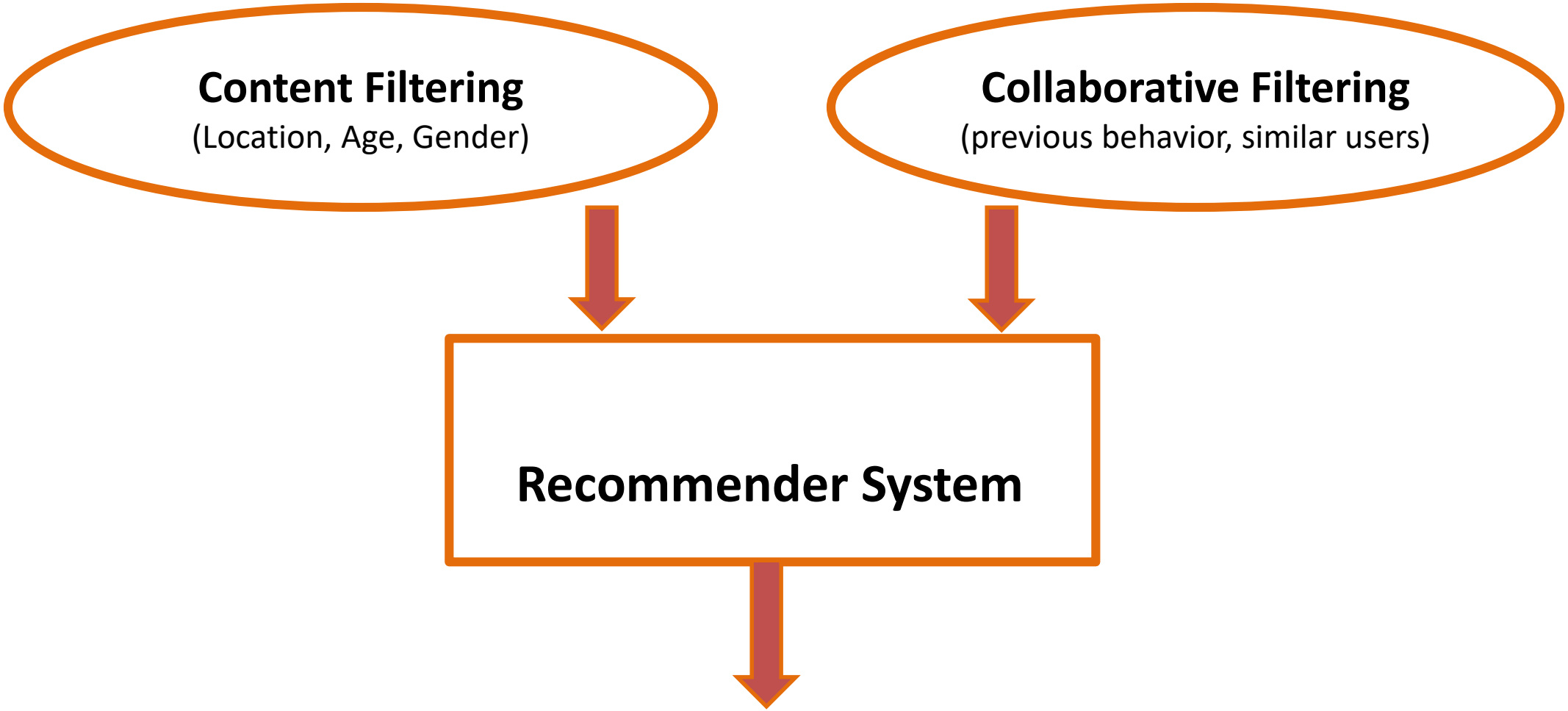
Basic Models of Recommender Systems

- The basic models for recommender systems work with two kinds of data, which are

- i. User-Item Interactions, such as ratings or buying behavior, and
- ii. The attribute information about the users and items such as textual profiles or relevant keywords.

- Content-based systems also use the ratings matrices in most cases, although the model is usually focused on the ratings of a single user rather than those of all users.

Types of Recommendation system



Basic Models of Recommender Systems

- In knowledge-based recommender systems, the recommendations are based on explicitly specified user requirements.

- Hybrid systems combine the strengths of various types of recommender systems to create techniques that can perform more robustly in a wide variety of settings.

Collaborative Filtering Models

- Collaborative filtering models use the collaborative power of the ratings by multiple users to make recommendations.

- The main challenge in designing collaborative filtering methods is that the underlying ratings matrices are sparse.

  Eg. Movie Recommendations.

Collaborative Filtering Models

- The basic idea of collaborative filtering methods is that these unspecified ratings can be imputed.

- Here the observed ratings are highly correlated across various users and items.

- Most of the models for collaborative filtering focus on leveraging either inter-item correlations or inter-user correlations for the prediction process. Some models also use both types of correlations.

Collaborative Filtering Models contd.

- There are two types of methods that are commonly used in collaborative filtering:

  a) Memory- based methods: they are also referred to as neighborhood-based collaborative filtering algorithms. In which the ratings of user-item combinations are predicted on the basis of their neighborhoods.

  These neighborhoods can be defined in one of two ways -

  i) User-based Collaborative filtering: The ratings provided by the like-minded users of a target user A are used in order to make the recommendations for A

- ii) Item-based collaborative filtering: To make the rating predictions for target item B by user A, the first step is to determine a set S of items that are most similar to target item B.

- The advantages of memory-based techniques are that they are simple to implement and the resulting recommendations are often easy to explain

Collaborative Filtering Models contd.

b) Model-based Methods: Here the machine learning and data mining methods are used in the context of predictive models.

In cases where the model is parameterized, the parameters of this model are learned within the context of an optimization framework.

- Examples: Decision trees, Rule-based models, Bayesian methods and latent factor models.

Collaborative Filtering Models contd.

- Collaborative filtering models are closely related to missing value analysis.
- It can be viewed as a special case of problems in which the data matrix is very large and sparse.
- It can also be viewed as generalizations of classification and regression modeling, here the class/dependent variable can be viewed as an attribute with missing values, other columns are treated as features/independent variables.

Neighborhood-Based Collaborative Filtering

- Neighborhood-based filtering algorithms can be formulated in one of two ways:

1. Predicting the rating value of a user-item combination: In this case, the missing rating r_{uj} of the user u for item j is predicted.

2. Determining the top-k items or top-k users: The problem of determine the top-k items is more common than that of finding the top-k users.

Neighborhood-Based Collaborative Filtering

- Item-based methods provide more relevant recommendations because of the fact that a user's own ratings are used to perform the recommendation.
- In item-based methods, similar items are identified to a target item, and the user's own ratings on those items are used to extrapolate the ratings of the target.
- Although item-based recommendations are often more likely to be accurate, the relative accuracy between item-based and user-based methods also depends on the data set at hand.

Collaborative Filtering

- Consider user x
- Find set N of other users whose ratings are "similar" to x's ratings
- Estimate x's ratings based on ratings of users in N.

DATA ANALYTICS

Similar Users(1):

	HP1	HP2	HP3	KGF	BB1	BB2	BB3
A	4			5	1		
B	5	5	4				
C				2	4	5	
D		3					3

- Consider users X and Y with rating vectors r_x and r_y
- We need a similarity metric sim(x, y)
- Capture intuition that sim(A,B) ?? sim (A,C)

DATA ANALYTICS

Option 1:Jaccard Similarity:

	HP1	HP2	HP3	KGF	BB1	BB2	BB3
A	4			5	1		
B	5	5	4				
C				2	4	5	
D		3					3

- $Sim(A, B) = |r_A \cap R_B| / |r_A \cup r_B|$
- Sim(A, B)<sim(A, C)
- Problem: Ignores rating values!

Option 2: Cosine similarity

	HP1	HP2	HP3	KGF	BB1	BB2	BB3
A	4	0	0	5	1	0	0
B	5	5	4	0	0	0	0
C	0	0	0	2	4	5	
D	0	3	0	0	0	0	3

- Sim(A, B) = $\cos(r_A, r_B)$
- Sim(A, B) = 0.38, Sim(A, C)= 0.32
- Sim(A,B)< sim(A,C), but not by much

- Problem: treats missing ratings as negative

Option 3: Centered cosine

- Normalize ratings by subtracting row mean

	HP1	HP2	HP3	KGF	BB1	BB2	BB3
A	4			5	1		
B	5	5	4				
C				2	4	5	
D		3					3

	HP1	HP2	HP3	KGF	BB1	BB2	BB3
A	4-10/3=2/3			5/3	-7/3		
B	1/3	1/3	-2/3				
C				-5/3	1/3	4/3	
D		0					0

Option 3: Centered cosine similarity(2)

	HP1	HP2	HP3	KGF	BB1	BB2	BB3
A	2/3			5/3	-7/3		
B	1/3	1/3	-2/3				
C				-5/3	1/3	4/3	
D		0					0

- Sim(A, B) = cos($r_{A,}$ r_B) =0.09; Sim(A, C) = -0.56

- Sim(A, B) > sim(A, C)

- Captures intuition better
- Missing ratings treated as “average”
- Handles “tough raters” and “ easy raters”
- Also known as Pearson Correlation

DATA ANALYTICS

Item-Item Collaborative Filtering:

- So far: User-user Collaborative filtering

- Another view: Item-Item
- For item I, find other similar items
- Estimate rating for item I based on ratings for similar items
- Can use same similarity metrics and prediction functions as in user-user model.

$$r_{xi} = \frac{\sum_{j \in N(i;x)} s_{ij} \cdot r_{xj}}{\sum_{j \in N(i;x)} s_{ij}}$$

- S_{ij} … Similarity of items I and j
- R_{xj}….Rating of user x on item j
- N(I;x)… set items rated by x similar to i

DATA ANALYTICS

Item-Item CF(|N|=2)

Users

Movies

	1	2	3	4	5	6	7	8	9	10	11	
1	1		3		?	5			5		4	
2			5	4			4			2	1	3
3	2	4		1	2		3		4	3	5	
4		2	4		5			4			2	
5			4	3	4	2					2	5
6	1		3		3			2			4	

Unknown Rating

Rating between 1to 5

- Estimate rating of movie 1 by user 5

DATA ANALYTICS

Item-Item CF(|N|=2)

Users

Movies	1	2	3	4	5	6	7	8	9	10	11		Sim(1,m)
1	1		3		?	5			5		4		1.00
2			5	4			4			2	1	3	-0.18
3	2	4		1	2		3		4	3	5		0.41
4		2	4		5			4			2		-0.10
5			4	3	4	2					2	5	-0.31
6	1		3		3			2			4		0.59

Here we use Pearson correlation as similarity

1) Subtract mean rating m from each movie I

M = (1+3+5+5+4)/5 = 3.6

Row 1:[-2.6, 0, -0.6, 0, 0, 1.4, 0, 0, 1.4, 0, 0, 0.4, 0]

2) Compute cosine similarities between rows

DATA ANALYTICS

Item-Item CF(|N|=2)

Users

Movies	1	2	3	4	5	6	7	8	9	10	11	
1	1		3		?	5			5		4	
2			5	4			4			2	1	3
3	2	4		1	2		3		4	3	5	
4		2	4		5			4			2	
5			4	3	4	2					2	5
6	1		3		3			2			4	

Predict by taking weighted average

r_{15} = (0.41*2 + 0.59*3) / (0.41 + 0.59) = 2.6

Item-Item vs. User-User

1. In theory, user-user and item-item are dual approaches

2. In practice, item-item outperforms user-user in many use cases.

3. Items are “simpler” than users

- items belong to a small set of “genres”, users have varied tastes.
- Item similarity is more meaningful than user similarity

References

"Recommender Systems, The Textbook by Charu C. Aggarwal, Springer 2016 – Sections 1.4 - 2.3.6

DATA ANALYTICS

Image Courtesy

http://www.mmds.org/mmds/v2.1/ch09-recsys1.pptx

https://www.researchgate.net/publication/287952023_Collaborative_Filtering_Recommender_Systems

http://cs229.stanford.edu/proj2014/Rahul%20Makhijani,%20Saleh%20Samaneh,%20Megh%20Mehta,%20Collaborative%20Filtering%20Recommender%20Systems.pdf

https://www.scribd.com/presentation/414445910/CS548S15-Showcase-Web-Mining

https://towardsdatascience.com/image-recommendation-engine-leverage-transfert-learning-ec9af32f5239

http://elico.rapid-i.com/recommender-extension.html
https://www.youtube.com/watch?v=h9gpufJFF-0

THANK YOU

Jyothi R.
Assistant Professor,
Department of Computer Science
jyothir@pes.edu