Collaborative Filtering (CF)

- What is collaborative filtering?
- Basic concepts and mechanism
- Examples
- Why CF?
- CF Algorithms
- User-User Similarity
- Item-Item Similarity
- Summary

1

Definition

- Filtering is a process of finding the most valuable and interesting information
- CF is a type of filtering, which employs other people information
- CF is recent technique for recommendation
- Relatively new concept & very popular
- Has many important applications in
  - E-commerce, search engines
  - Direct recommendations (books, movies, etc.)
- Used to cope with information overload
- With the growth of e-commerce, it is becoming widely used technique by online vendors

2

Basic Concepts

- Goal: to predict the preferences of an active user based on the preferences of other users
- Idea: active user prefers those items that like-minded users prefer or dissimilar users do not
- Assumption: if users U_1 and U_2 rate j_u items similarly, they share similar tastes, and hence will rate other items similarly
- Tasks:
  - Prediction: referrals for single items
  - Top-*N* Recommendation: sorted item list

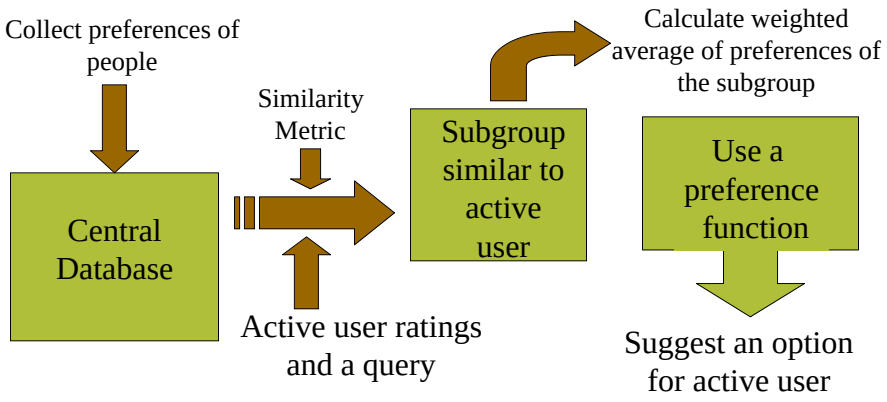
3

Basic Concepts

- If users A and B rate k items similarly, they share similar tastes, and hence they will rate other items similarly
- CF approaches differ in
  - How they define a “rating”
  - How they define “k”
  - How they define “similarly”

4

Basic Mechanism



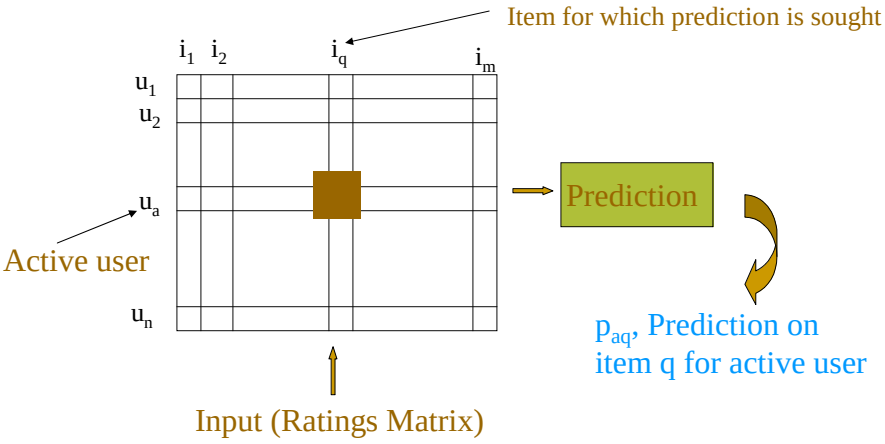
5

Basic Mechanism

- A large group of people's preferences are collected
- Using a similarity metric, a subgroup of people is selected whose preferences are similar to the preferences of the person who seeks advice
- Weighted average of the preferences for that subgroup is calculated
- Prediction formula is used to find prediction for the person who seeks advice

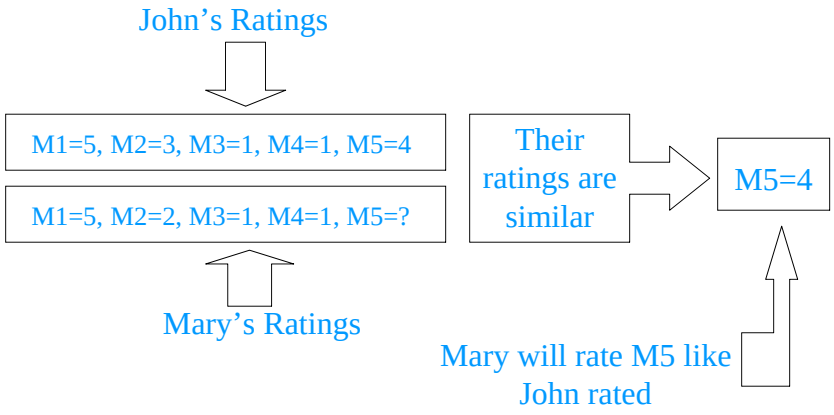
6

Collaborative Filtering Process



7

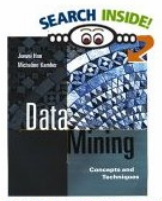
An Example



8

Example: Recommendation

Data Mining: Concepts and Techniques
by Jiawei Han (Author), Micheline Kamber (Author)



Search inside this book

List Price: $54.95
Price: **$54.95** & This item ships for **FREE with Super Saver Shipping**. See details.
Availability: Usually ships within 24 hours

Want it delivered tomorrow, March 4? Order it in the next 1 hour and 57 minutes, and choose **One-Day Shipping** at checkout. See details.

8 used & new from $44.95

Edition: Hardcover

Customers who bought this book also bought:

- Data Preparation for Data Mining: by Dorian Pyle (Author)
- The Elements of Statistical Learning: by T. Hastie, et al
- Data Mining: Introductory and Advanced Topics: by Margaret H. Dunham
- Mining the Web: Analysis of Hypertext and Semi Structured Data

9

Why CF?

Problem: ***Information Overload***

Solution: ***Collaborative Filtering (CF)***

10

Why CF?

- Information overload is becoming a problem
- With the growth of e-commerce, products to buy are increasing
- Customers want to buy what they like without wasting their time
- Online vendors want to keep their customers
- Lots of online products
- Reduce choices

11

CF Algorithms

- Memory-based: operate over entire user database
- Model-based: uses user database to estimate a model, then uses that model for predictions
- Hybrid: combine memory and model based algorithms

12

CF Algorithms

- Collaborative filtering algorithms
  - Information Tapestry
  - GroupLens
  - Ringo Music Recommender
  - Bellcore Video Recommender
  - PHOAKS, Referral Web, and the Fab System
  - SVD-based CF, Eigentaste
  - CF with naïve Bayesian classifier
  - Jester 2.0, SWAMI, CF with Personality Diagnosis

13

Ratings

- Each user has a profile
- Users rate items
  - Explicitly: score from 1..5
  - Implicitly: web usage mining
    - Time spent in viewing the item
    - Navigation path
    - Etc…
- Ratings
  - Binary
  - Numerical

14

Basic Approaches

- Collaborative Filtering (CF)
  - Look at users collective behavior
  - Look at the active user history
  - Combine!
- Content-based Filtering
  - Recommend items based on key-words
  - More appropriate for information retrieval

15

CF Parts

- CF has two parts:
  - Filtering part: guiding people's choices of what to read, what to look at, what to watch, and what to listen to
  - Collaborative part: doing that guidance based on information gathered from some other people

16

Collaborative Filtering: A Framework

Items: I

	i_1	i_2	…	i_j	…	i_n
u_1	3	1.5	…. …			2
u_2						
…	2					
u_i	1			r_{ij}=?		
…						
u_m	3					

Users: U

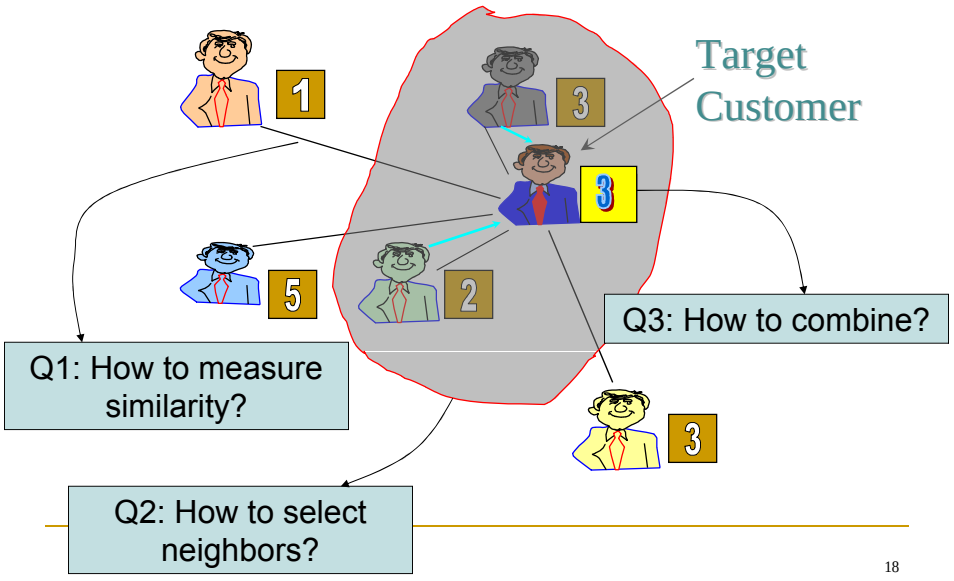
The task:

Q1: Find unknown ratings?

Q2: Which items should we recommend to this user?

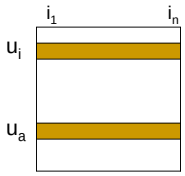
17

User-User Similarity: Intuition



18

How to Measure Similarity?



- Pearson correlation coefficient

$$w_p(a,i) = \frac{\sum_{j \in \text{Commonly Rated Items}} (r_{aj} - \bar{r}_a)(r_{ij} - \bar{r}_i)}{\sqrt{\sum_{j \in \text{Commonly Rated Items}} (r_{aj} - \bar{r}_a)^2 \sum_{j \in \text{Commonly Rated Items}} (r_{ij} - \bar{r}_i)^2}}$$

- Cosine measure
  - Users are vectors in product-dimension space

$$w_c(a,i) = \frac{r_a . r_i}{\|r_a\|_2 * \|r_i\|_2}$$

19

Nearest Neighbor Approaches

- Offline phase:
  - Do nothing…just store transactions
- Online phase:
  - Identify highly similar users to the active one
    - Best K ones
    - All with a measure greater than a threshold
- Prediction

$$r_{aj} = \bar{r}_a + \frac{\sum_i w(a,i)(r_{ij} - \bar{r}_i)}{\sum_i w(a,i)}$$

User a's neutral

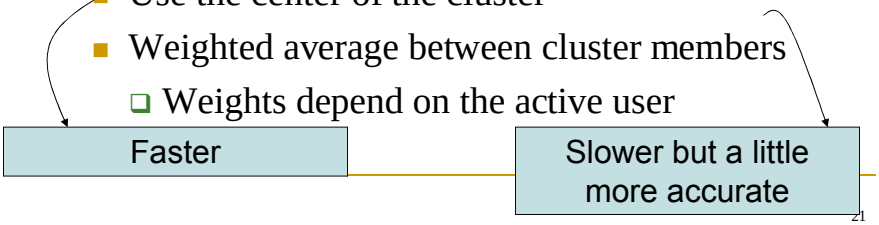
User i's deviation

User a's estimated deviation

20

Clustering

- Offline phase:
  - Build clusters: k-mean, k-medoid, etc.
- Online phase:
  - Identify the nearest cluster to the active user
  - Prediction:
    - Use the center of the cluster
    - Weighted average between cluster members
      - Weights depend on the active user



21

Limitations of CF

- Problems:
  - Sparsity
  - Scalability
  - Synonymy
- Goal:
  - Accurate
  - Efficient referrals

22

How to Measure Similarity?

- Q1:How to measure similarity?

Done… Really??

Sparsity results from the poor representation!

$$w_p(a,i) = \frac{\sum_{j \in \text{Commonly Rated Items}} \ldots\ldots}{\ldots\ldots}$$

U1 rates *recycled letter pads* High
U2 rates *recycled memo pads* High

Both of them like *Recycled office products*

They are similar but the math won't work for that

What about Sparsity?
Not enough common Items
implies spurious neighbors
and hence bad recommendations

By working at the right level of abstraction we can eliminate sparsity

User-User Methods Evaluation

- Achieve good quality in practice
- The more processing we push offline, the better the method scale
- However:
  - User preference is dynamic
    - High update frequency of offline-calculated information
  - No recommendation for new users
    - We don't know much about them yet

24

Item-Item Similarity: The Intuition

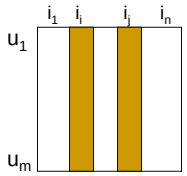
- Search for similarities among items
- All computations can be done offline
- Item-Item similarity is more stable than user-user similarity
  - No need for frequent updates
- Correlation Analysis
- Linear Regression

25

Correlation-based Methods

- Same as in user-user similarity but on item vectors
- Pearson correlation coefficient
  - Look for users who rated both items

$$s_{ij} = \frac{\sum_{u \in \text{Users Rated Both Items}} (r_{uj} - \bar{r}_j)(r_{ui} - \bar{r}_i)}{\sqrt{\sum_{u \in \text{Users Rated Both Items}} (r_{uj} - \bar{r}_j)^2 \sum_{u \in \text{Users Rated Both Items}} (r_{ui} - \bar{r}_i)^2}}$$

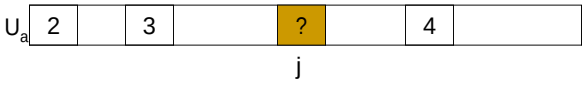


26

Correlation-based Methods

- Offline phase:
  - Calculate n(n-1) similarity measures
  - For each item
    - Determine its k-most similar items
- Online phase:
  - Predict rating for a given user-item pair as a weighted sum over similar items that he rated

$$r_{aj} = \frac{\sum_{i \in similar\ \ items} s_{ij}\ r_{ai}}{\sum_{i \in similar\ \ items} s_{ij}}$$



27

Summary

- CF is widely used by online vendors
- It has important applications
- Many CF algorithms
- CF tasks: Predictions and top-N recommendations
- User-user & item-item methods
- Memory- or model-based approaches
- CF has some disadvantages
  - Threat to individual privacy

28