
Information Retrieval

Lecture 10 - December 19th, 2007

Recommender Systems & Social Filtering



Thomas Hofmann
thofmann@google.com

Overview

1. Recommender Systems
2. Neighbor-based Collaborative Filtering
3. Evaluation Metrics
4. Probabilistic Model (pLSA)

Recommender Systems

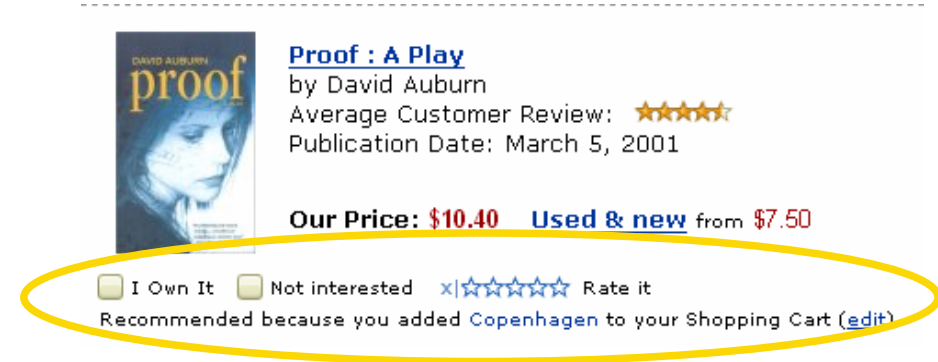
- **Recommender systems** are designed to automatically recommend items (e.g. products, documents, etc.) to users in a **personalized** way - matchmaking between users and items
- Comparison with search engines
 - search or ad hoc retrieval: **query** expresses “one-time” information need
 - recommender systems: query-free based on **user profile**
- Application areas
 - E-commerce (amazon.com, cdnow.com, etc.)
 - Information retrieval and agent systems (“push” technologies)
 - Intelligent user interfaces
 - ...

The logo for amazon.com, featuring the text "amazon.com" in a sans-serif font with a curved arrow underneath.The logo for CDNOW, consisting of the text "CDNOW" in white inside a dark blue rounded rectangle, with the tagline "Never miss a beat." below it.The logo for moodlogic, with the text "moodlogic" in a stylized font where "mood" is blue and "logic" is yellow, set against a blue background.The logo for movielens, with the text "movielens" in red and the tagline "helping you find the right movies" below it.

Amazon.com

User Profiles

- **User profiles** may store various data related to user interaction history
- **explicit feedback:**
 - user ratings
 - reviews
 - auctions
 - questionnaires
- **implicit feedback:**
 - page visits
 - purchase history
 - download log
 - click-stream data & browsing paths
 - time spend reading



Histories

Different time scales for historic data

- Long-term interest profiles
 - e.g. complete customer history
 - captures persistent interest
- Short-term interest profiles
 - e.g. session-level history
 - captures specific tasks and focus of attention
 - [seems to work well in e-commerce applications]

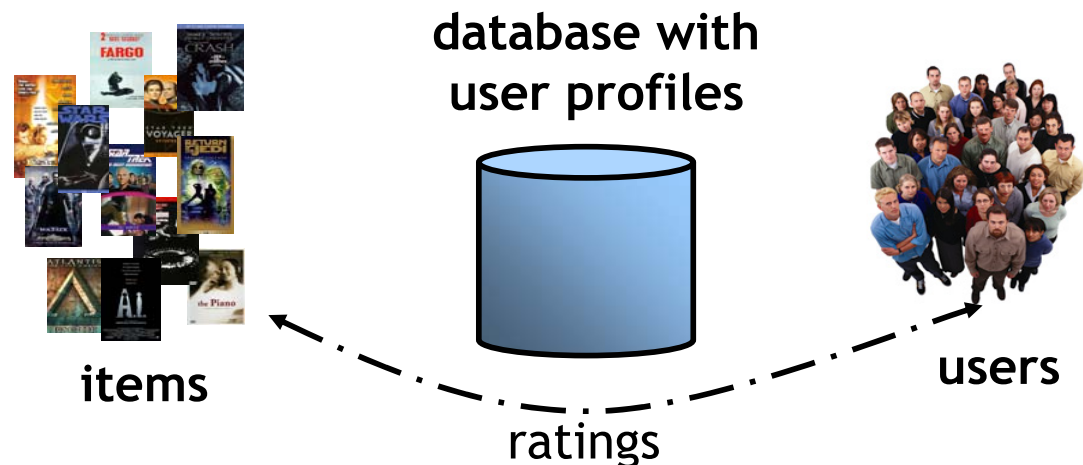
Collaborative & Content-based Filtering

- **Content-based filtering**

- properties of objects or similarities between objects are used to improve predictions
 - examples: movie - cast, director, genre; document - words & phrases; product - category, price, etc.

- **Collaborative/social filtering**

- properties of persons or **similarities between persons** are used to improve predictions.
- makes use of user profile data

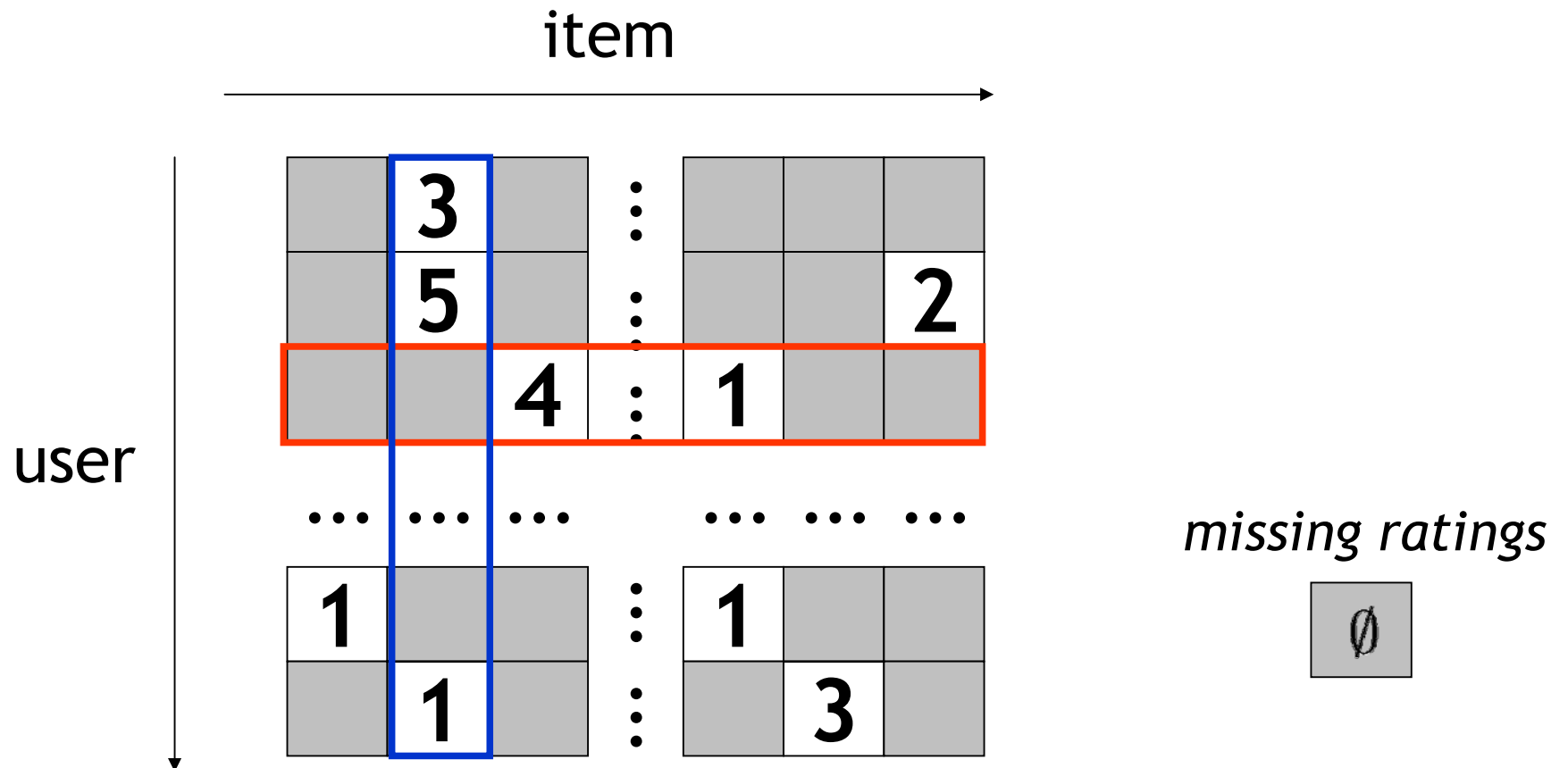


Advantages of Collaborative Filtering

- Support for items whos **content is not easily analyzed**
 - e.g. multimedia objects
- Filter items based on **quality and taste**
 - not just based on content
- Ability to provide **serendipitous recommendations**
 - “: the faculty or phenomenon of finding valuable or agreeable things not sought for” (Merian Webster)

Rating Matrix

Rating matrix is typically a large matrix with many (mostly) **missing values**



Example: User-Item Rating Matrix

	Matrix Reloaded	Lord of the Rings 2	Titanic	La vita è bella
Alice	2	5		5
Bob	5		1	3
Carol		5		
Dean	2	5	5	4

Approaches & Methods

- Neighborhood-based or **memory-based** methods
 - **active user**: user for which to make a recommendation
 - select a subset of appropriate, i.e. similar users
 - compute a weighted aggregate of their ratings
- **Model-based** methods
 - build a (statistical) model over rating matrix
 - predict missing ratings by inference

Neighborhood-based Methods

Typical four step approach

1. Weight all users w.r.t. their similarity with the active user
2. Select a subset of users to use as a set of predictors
3. Individually normalize selected user ratings
4. Compute a prediction from a weighted combination of selected neighbors' normalized rating

Notation & Definitions

i, j : generic user indices

k, l : generic item indices

\emptyset : unobserved rating

v_{ik} : rating of user i on item k

I_i : set of items a user has rated, $I_i \equiv \{k : v_{ik} \neq \emptyset\}$

n_i : number of observed ratings for user i , $n_i \equiv |I_i|$

\bar{v}_i : mean rating of user i , $\bar{v}_i \equiv \frac{\sum_{k \in I_i} v_{ik}}{n_i}$

σ_i : standard deviation of ratings of user i , $\sigma_i \equiv \frac{\sum_{k \in I_i} (v_{ik} - \bar{v}_i)^2}{n_i}$

Pearson Correlation Coefficient

Similarity between user can be defined via statistical correlation measure

$$s_{ij} \equiv \frac{\sum_{k \in I_i \cap I_j} (v_{ik} - \bar{v}_i)(v_{jk} - \bar{v}_j)}{\sigma_i \sigma_j}$$

- Ratings are mean-normalized:
 - accounts for the fact that some users are more positive/negative with their ratings
 - calibrates a rating w.r.t. the users “scale”
- Variance normalization
 - takes into account “dynamic range” of ratings
- Completely observed case:
 - inner product of so-called z-scores (mean zero, unit variance)
- Notice: correlation can also be negative (anti-correlated users)

GroupLens Algorithm

Weighted prediction based on Pearson correlation similarity

$$\hat{v}_{ik} = \bar{v}_i + \frac{\sum_{j:k \in I_j} s_{ij} (v_{jk} - \bar{v}_j)}{\sum_{j:k \in I_j} |s_{ij}|}$$

- Mean rating is modeled separately
- Correction relative to mean is predicted based on similar users
- Only users who have rated an item contribute to a prediction for that item
- User contributions are weighted by their similarity
- Prediction is normalized by sum of similarities

Variants & Tricks of the Trade

- Neighborhood pruning
 - only include R most similar users that have rated a particular item (e.g. R=40)
 - improves accuracy and reduces computational complexity
- Significance weighting
 - give more weights to users with more co-rated items
 - simple heuristics:
 - if $n = \# \text{co-rated items} < 50$, then down-weight by $n/50$

Item-based Recommendation

- Swap role of users and items
- Compute similarity between items based on a correlation measure (e.g. Pearson)
- Empirically yields often better results than user-based recommendation

Metrics

Mean absolute prediction error **MAE** (over some test set T)

$$MAE(T) = \frac{1}{|T|} \sum_{(i,k) \in T} |\hat{v}_{ik} - v_{ik}|$$

(alternatively: average for each user first and compute the mean)

Recommendation ranking

$$R_i = \sum_{k: \text{ordered}} 2^{-\rho(k-1)} \max(v_{ik} - d, 0), \quad R = 100 \frac{\sum_i R_i}{\sum R_i^{\max}}$$

ρ : viewing half-live (e.g. 5), d : neutral vote, R_{\max} : maximal achievable utility

Experimental Protocols

Leave-out protocol

- **All-but-1** protocol:
 - leave out one rating for every user
 - predict missing rating - or - determine rank
- All-but-K protocol:
 - leave out K ratings per user

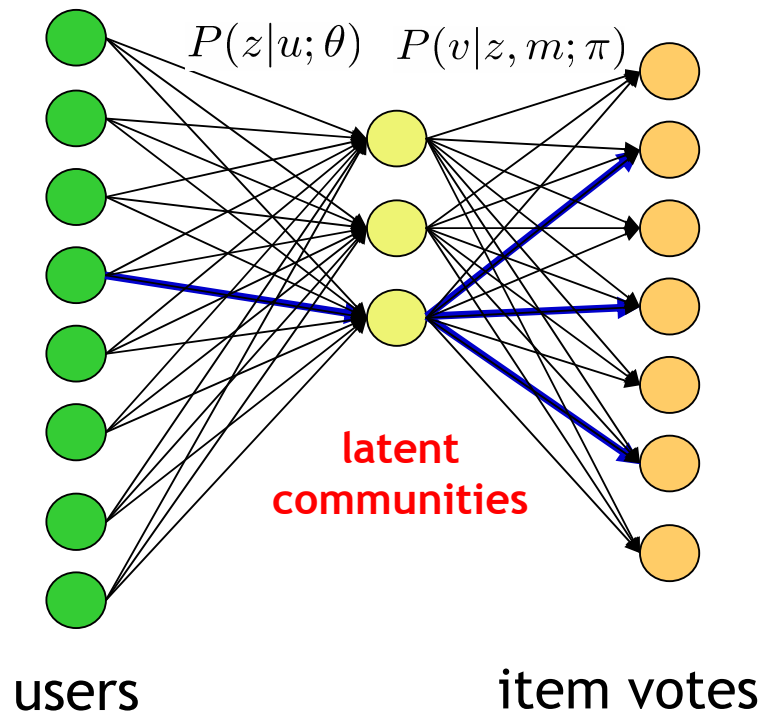
Given K protocol

- samples K ratings for each user (that has at least K+1) ratings
- test using remaining ratings
- allows investigating “cold-start” effect and influence of number of available ratings on accuracy

Datasets

- ▶ EachMovie
 - ▶ collected by HP/Compaq Research (formerly DEC Research)
 - ▶ 72916 users
 - ▶ 1628 movies
 - ▶ 2811983 ratings
- ▶ MovieLens
 - ▶ collected by GroupLens project
- ▶ Jester database: jokes
 - ▶ 4.1 million continuous ratings (-10.00 to +10.00) of 100 jokes from 73,496 users
- ▶ BookCrossing (BX) dataset
 - ▶ 278,858 users (anonymized but with demographic information) providing 1,149,780 ratings (explicit / implicit) about 271,379 books
- ▶ Download site: <http://www.cs.umn.edu/Research/GroupLens/>

Social Filtering via pLSA



$$P(\mathbf{v}|u, m) = \sum_z P(\mathbf{v}|z, m)P(z|u)$$

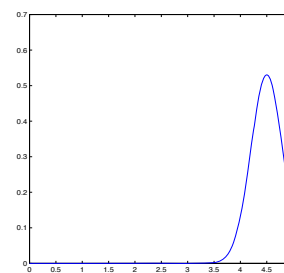
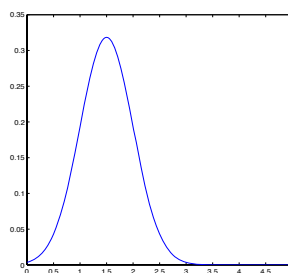
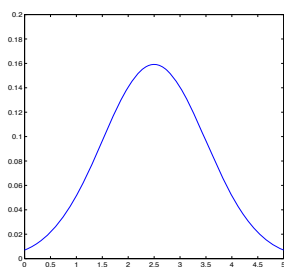
Diagram illustrating the equation above, with arrows pointing from the variables to their corresponding labels in boxes below:

- u points to **user**
- m points to **item**
- $P(\mathbf{v}|z, m)$ points to **Gaussian or multinomial distribution**

- every person is characterized by a distribution over interest groups
- rating is independent of person, given state of latent variable (community)
- analogy to information retrieval: person=document, item=word

Gaussian Probabilistic LSA

Illustration: characteristics of a user community



$$P(v|'Matrix', z) \quad P(v|'Atlantis', z) \quad P(v|'Existenz', z)$$

Interests Groups, Eachmovie

D:\inout\Visual.html - Microsoft Internet Explorer

File Edit View Favorites Tools Help Links >>

Interest Group 1, *4.8*	Interest Group 2, *4.6*	Interest Group 3, *4.5*	Interest Group 4, *4.4*	Interest Group 5, *4.4*
Twister [4.6*] [0.064]	Batman (1989) [4.1*] [0.066]	Trainspotting [5*] [0.038]	Dead Man Walking [5*] [0.052]	The Santa Clause [4.5*] [0.014]
Independence Day (...) [4.9*] [0.061]	Apollo 13 [5*] [0.065]	Fargo [5*] [0.033]	The Truth about Ca... [4.3*] [0.039]	Casper [4.5*] [0.014]
Toy Story [4.9*] [0.057]	True Lies [4.7*] [0.059]	Pulp Fiction [5*] [0.028]	Get Shorty [4.6*] [0.036]	Robin Hood: Men in... [4.3*] [0.013]
Broken Arrow [4.4*] [0.054]	Batman Forever [4.1*] [0.054]	Clerks [4.7*] [0.023]	Sense and Sensibil... [5*] [0.035]	Tommy Boy [4.5*] [0.013]
Interest Group 6, *4.3*	Interest Group 7, *4.3*	Interest Group 8, *4.2*	Interest Group 9, *4*	Interest Group 10, *3.9*
The Remains of the... [4.5*] [0.047]	The Empire Strikes... [4.7*] [0.032]	Pretty Woman [4.3*] [0.059]	Sleepers [4.2*] [0.015]	A Clockwork Orange... [4.2*] [0.01]
The Piano [4.7*] [0.043]	Raiders of the Los... [4.7*] [0.03]	Mrs. Doubtfire [4.3*] [0.059]	Jemmy Maguire [4.6*] [0.013]	Amadeus (1984) [4.2*] [0.0098]
Like Water For Cho... [4.7*] [0.043]	Star Wars [4.9*] [0.026]	Ghost [4.4*] [0.057]	The First Wives Cl... [3.8*] [0.013]	Psycho (1960) [4.3*] [0.0098]
Much Ado About Not... [4.6*] [0.041]	Indiana Jones and ... [4.5*] [0.025]	Sleepless in Seatt... [4.4*] [0.055]	William Shakespear... [4.5*] [0.011]	One Flew Over the ... [4.5*] [0.0095]

Des-interests Groups, Eachmovie

Interest Group 31, *2.2*	Interest Group 32, *2*	Interest Group 33, *1.8*	Interest Group 34, *1.8*	Interest Group 35, *1.7*
E.T.: The Extrater... [^{2.6*} [0.01]	Lord of Illusions [^{1.6*} [0.011]	Sleepless in Seatt... [^{1.8*} [0.017]	Toy Story [^{2.4*} [0.05]	Striptease [^{0.025*} [0.033]
The Sound of Music... [^{2.3*} [0.0086]	Tales From the Hoo... [^{1.6*} [0.0087]	The Firm [^{1.8*} [0.015]	Mission: Impossible... [^{1.8*} [0.049]	Independence Day (...) [^{0.87*} [0.029]
Top Gun (1986) [^{2.3*} [0.0086]	Mallrats [^{2.4*} [0.0083]	Pretty Woman [^{1.5*} [0.015]	Independence Day (...) [^{2.1*} [0.048]	The Cable Guy [^{0.16*} [0.028]
Mary Poppins (1964... [^{2.3*} [0.0083]	Wes Craven's New N... [^{2*} [0.0082]	Dave [^{2*} [0.015]	Twister [^{1.8*} [0.043]	Barb Wire [^{4.9e-005*} [0.025]
Interest Group 36, *1.1*	Interest Group 37, *0.68*	Interest Group 38, *0.39*	Interest Group 39, *0.16*	Interest Group 40, *0.16*
Super Mario Bros. [^{0.11*} [0.017]	Mighty Morphin Pow... [^{0.017*} [0.033]	Dumb and Dumber [^{0.0025*} [0.038]	Kazaam [^{0.028*} [0.014]	Tales From the Hoo... [^{0.022*} [0.0075]
The Beverly Hillbi... [^{0.34*} [0.016]	The Brady Bunch Mo... [^{0.28*} [0.024]	Ace Ventura: Pet D... [^{0.016*} [0.034]	Children of the Co... [^{0.021*} [0.014]	Vampire in Brookly... [^{1.3e-005*} [0.007]
Richie Rich [^{0.22*} [0.015]	Mortal Kombat [^{0.21*} [0.018]	Ace Ventura: When ... [^{0.00067*} [0.033]	A Very Brady Seque... [^{0.083*} [0.012]	The Baby-Sitters C... [^{0.0063*} [0.007]
The Next Karate Ki... [^{0.21*} [0.014]	The Bridges of Mad... [^{0.015*} [0.018]	Waterworld [^{0.034*} [0.028]	Halloween: The Cur... [^{0.035*} [0.012]	Candyman: Farewell... [^{0.0039*} [0.0065]

Model-based vs. Memory-based

- Memory requirements
 - Memory-based: complete data set required for making recommendations
 - Model-based: compact summarization of data in statistical model
- Simplicity
 - Memory-based: straightforward to implement, no learning stage necessary
 - Model-based: more sophisticated algorithms, require learning/model fitting stage
- Accuracy
 - Memory-based: overall good accuracy
 - Model-based: (sometimes/somewhat) higher accuracy
- Optimization
 - Memory-based: basically one-fits-all approach
 - Model-based: can be tailored to specific tasks and objectives
- Privacy
 - Memory-based: raw data is stored
 - Model-based: raw data is only needed for model building

Trends & Topics

- Combining content-based and collaborative filtering
- Trust-awareness in recommender systems
- Privacy preserving collaborative filtering
- Temporal modeling
- Decentralized data management