

Recommendation Systems

N.Ravitha Rajalakshmi
Assistant Professor, Dept of IT
PSG College of Technology

Recommendation system


“Recommendation Systems are software agents that elicit the **interests and preferences** of individual **consumers** and make **recommendations** accordingly.”

[Xiao & Benbasat, MISQ, 2007]

Recommender system application areas

Online Shopping

You may also like



Jack & Jones
JAME - Polo shirt - orange
£21.00
Free delivery & returns


ALTERNATIVE PRODUCTS

Beko Washing Machine
Code: WMB81431LW
£269.99


Zanussi Washing Machine
Code: ZWH6130P
£269.99

Blomberg Washing Machine
Code: WNF6221
£299.99


You may also like



★★★★☆ (109)




★★★★★ (53)




★★★★☆ (33)

Social Media


Groups You May Like [More »](#)



Advances in Preference Handling
[Join](#)






FP7 Information and Communication Technologies (ICT)
[Join](#)




The Blakemore Foundation
[Join](#)

Real time examples


YouTube   [Upload](#) 


[What to Watch](#) [My Subscriptions](#) [Music](#)


Recommended





Snow White and the Seven Dwarfs - Grimm's Fairy ...
by Cartoons for Kids
6,628,925 views • 1 year ago



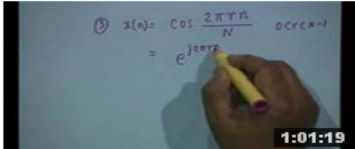
DIY Accessories: How to Make a Cute Bow Ring & ...
by Ann Le {Anneorshine} 
2,626,786 views • 1 year ago





Tangled Ever After: The Pursuit
by Disney Movie Trailers 
1,475,027 views • 3 years ago





Tangled - When Will My Life Begin - Mandy Moore
by sandeep2kd
22,794,508 views • 3 years ago



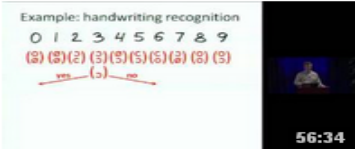
Lecture - 9 Discrete Fourier Transform (DFT)
by npTELhnd 
136,932 views • 6 years ago



Dark Energy & The Big Rip - Sixty Symbols
by Sixty Symbols 
219,696 views • 11 months ago



SURIYA, Jyothika UNSEEN RARE OFFICIAL MARRIAG...
by gvrchannel
222,742 views • 2 years ago



Machine Learning: The Basics, with Ron ...
by LinkedInTechTalks
53,868 views • 2 years ago

[Show more](#)

- Thousands of news articles and blog posts are created every day.
- Millions of movies, programs are streamed online.

Information
overload



Information Overload - Problem

- Lots of option creates more confusion.
- Problem –

“ How can the user find interesting information”

- **Recommendation** is widely used to mitigate the problem of information overload.

Benefits

Customer

- Narrows down the set of choices offered
- It helps the user to find interesting items.

Provider

- Increases the revenue.
- Increase user satisfaction.
- Sell more diverse items



- **Netflix** : 2/3 of the movies watched are recommended

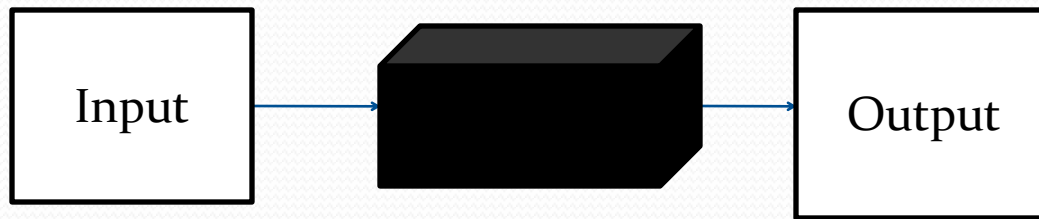
- **Google news**: recommendations generate 38% more click through.

- **Amazon**: 35% sales from recommendation



Recommendation systems

- Estimate a **function** that automatically **predicts** “how the user will like an item”



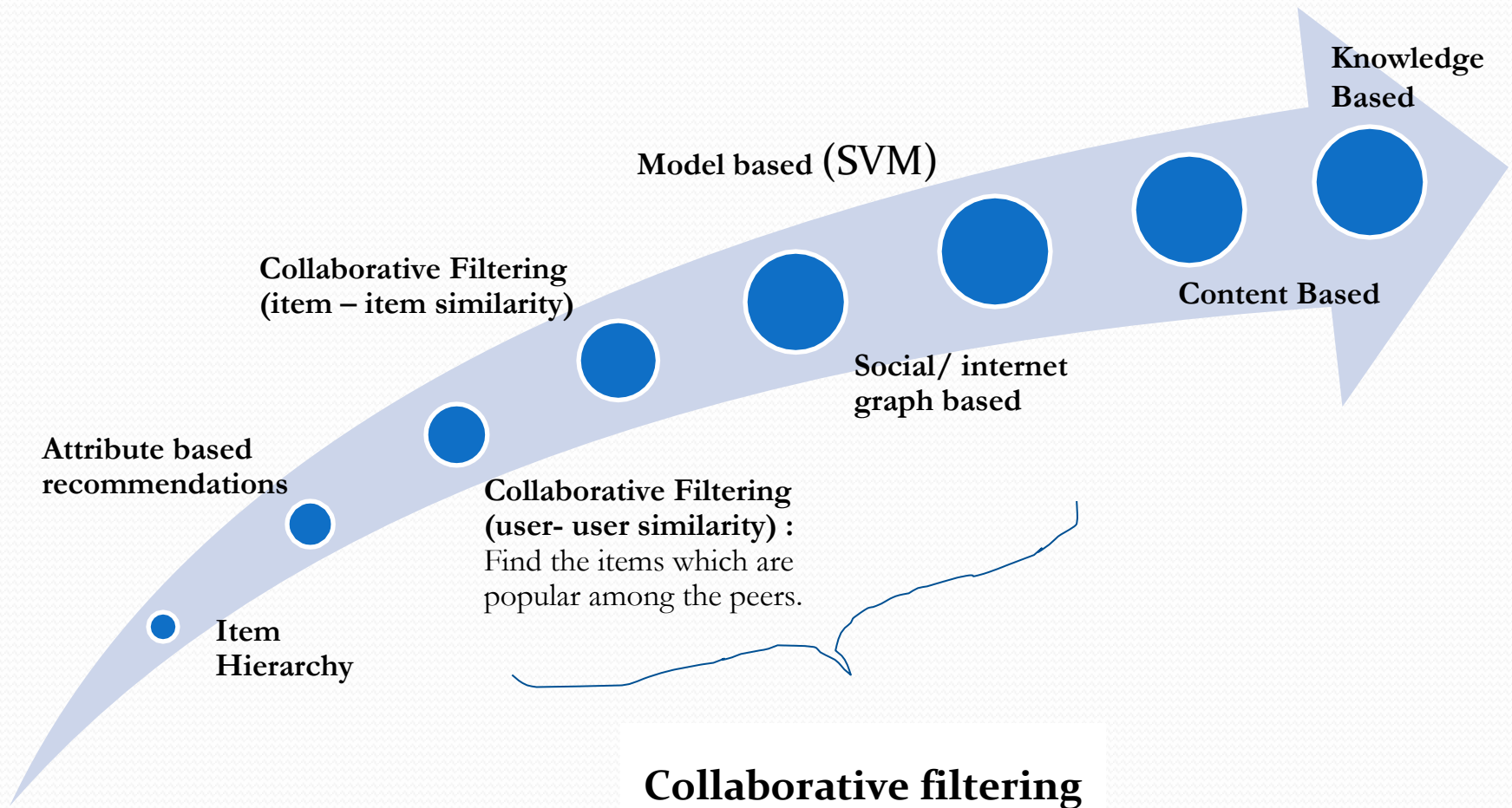
User profile parameters

- Past preferences
- Demographic characteristics

Relevant Items

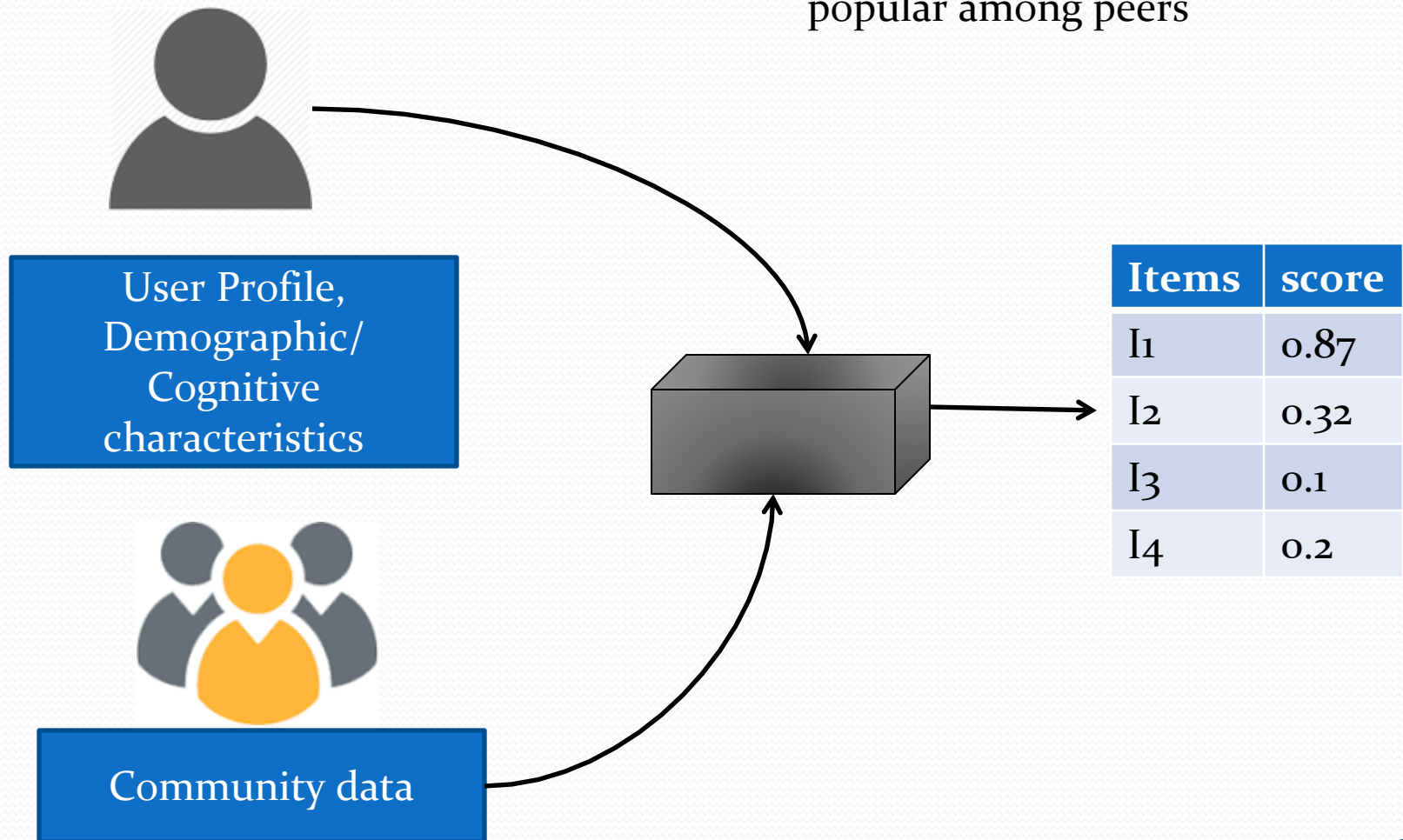
Items	score
I ₁	0.87
I ₂	0.32
I ₃	0.1
I ₄	0.2

Recommendation Approaches



Collaborative filtering

Find items which are popular among peers



Collaborative filtering

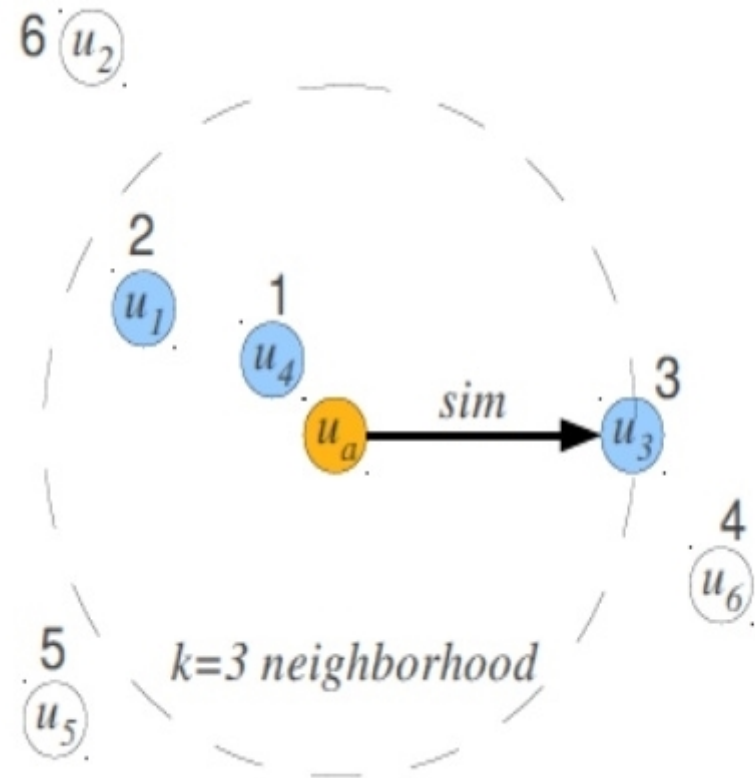
- Most Prominent approach to generate recommendations.
 - Used by large commercial, ecommerce sites.
 - Various state of art algorithms exist.
- Basic **assumptions**:
 - User provide ratings to the items either explicitly or implicitly.
 - These ratings are used in recommending the **highly relevant items** which are not yet rated by the user.

User based Collaborative Filtering

- It is used in Digg.
- Collaborative filtering makes use of ratings matrix.
- Basic steps:
 - Identify the users who are similar to the target/active user.
 - Identify the products liked by the similar users.
 - For each item not rated by the target user, predict the rating.
 - Recommend a set of top N products.

User based Collaborative Filtering

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8
u_1	?	4.0	4.0	2.0	1.0	2.0	?	?
u_2	3.0	?	?	?	5.0	1.0	?	?
u_3	3.0	?	?	3.0	2.0	2.0	?	3.0
u_4	4.0	?	?	2.0	1.0	1.0	2.0	4.0
u_5	1.0	1.0	?	?	?	?	?	1.0
u_6	?	1.0	?	?	1.0	1.0	?	1.0
u_a	?	?	4.0	3.0	?	1.0	?	5.0
r_a	3.5	4.0			1.3		2.0	



User based Collaborative Filtering

- How to measure similarity between users?
 - Popular similarity metric used is Pearson correlation coefficient. It takes value between -1 and $+1$.

$$sim(a, b) = \frac{\sum_{p \in P} (r_{a,p} - \bar{r}_a)(r_{b,p} - \bar{r}_b)}{\sqrt{\sum_{p \in P} (r_{a,p} - \bar{r}_a)^2} \sqrt{\sum_{p \in P} (r_{b,p} - \bar{r}_b)^2}}$$

a, b - vectors of users

P - list of items rated by both user a and b

\bar{r}_a - average rating of user a

Contd...

- How many neighbors to be considered?
 - It depends on the user. But number of neighbors would greatly influence the output.
- How to generate prediction from neighbor's ratings?
 - Common prediction function used:

$$pred(a, p) = \bar{r}_a + \frac{\sum_{b \in N} sim(a, b) * (r_{b,p} - \bar{r}_b)}{\sum_{b \in N} sim(a, b)}$$

Contd...

$$\text{sim}(1,4) = \frac{(4 - 3.33)(4 - 3) + (1 - 3.33)(2 - 3)}{\sqrt{(4 - 3.33)^2 + (1 - 3.33)^2} \sqrt{(4 - 3)^2 + (2 - 3)^2}}$$

Active user



I ₁	I ₂	I ₃	I ₄	I ₅	I ₆
5		4			1
		5		2	
	1		5		4
		4			2
4	5		1		

$$\text{sim}(1,4) = 0.88$$

NA

Contd...

$$Pred(4,1) = 3 + \frac{(0.88 * (5 - 3.33))}{0.88}$$



I1	I2	I3	I4	I5	I6
5		4			1
		5		2	
	1		5		4
4.6		4			2
4	5		1		

0.88

1

-1

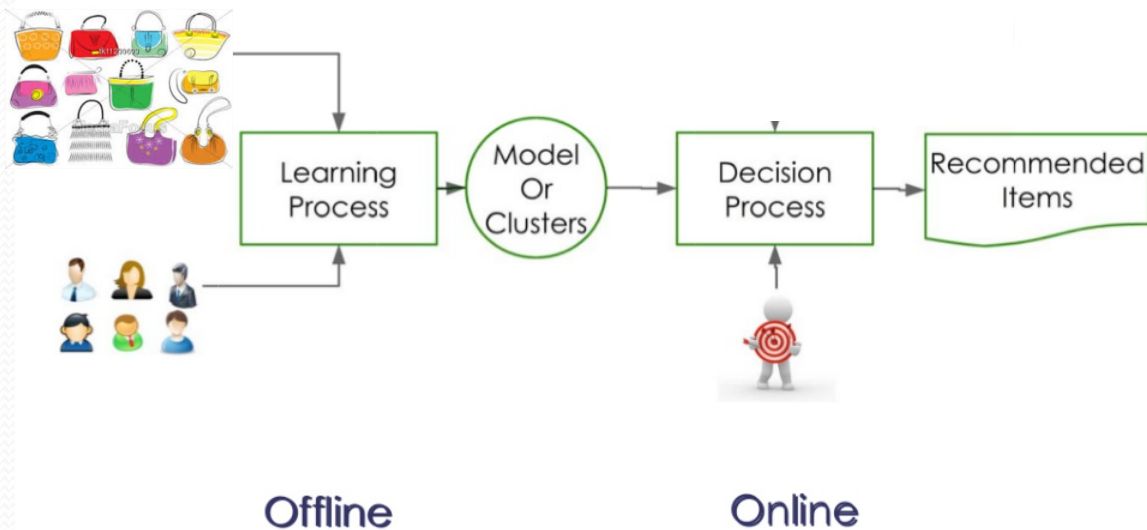
NA

Memory - based vs Model - based

- **User based** Collaborative Filtering is said to be **memory based**.
 - Straightforward and simple algorithm.
 - Provides good predictions.
 - Doesn't scale well for real time scenarios.
- Model based approaches:
 - Two step process
 - Model building / Training phase and Model usage phase
 - Many algorithms are used : SVD, Probabilistic models etc.
 - Model building phase is computationally expensive.

Model – based Approaches

Two-step process



Item based Collaborative Filtering

- Ex: Amazon, Netflix , YouTube etc.
- Basic idea:
 - It uses **similarity** between the **items** based on the ratings given by the user.

	Item1	Item2	Item3	Item4	Item5
Alice	5	3	4	4	?
User1	3	1	2	3	3
User2	4	3	4	3	5
User3	3	3	1	5	4
User4	1	5	5	2	1

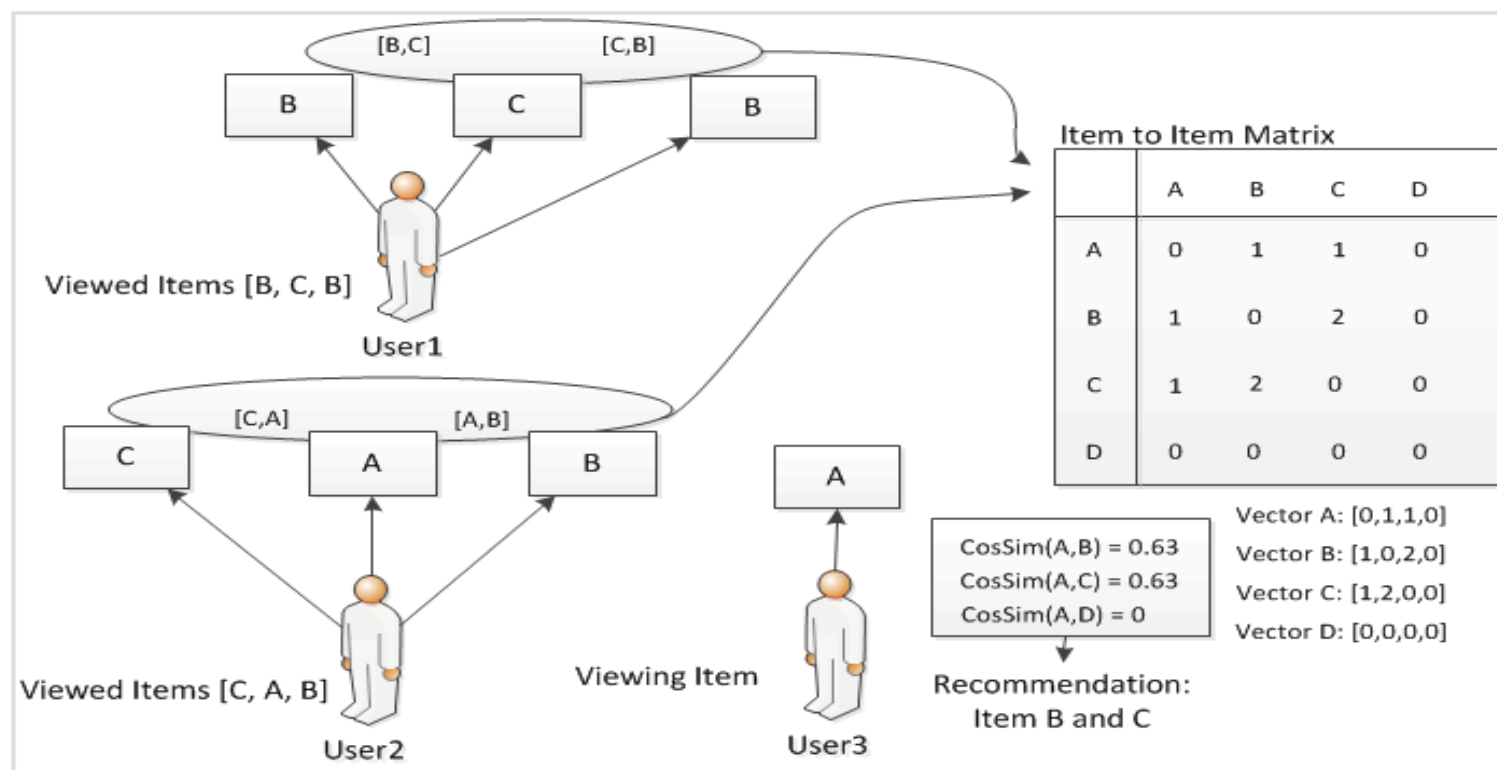
Contd...

- Ratings are considered as n-dimensional vector.
 - N-represents the total no of users.
- Similarity is calculated based on the angle between the vectors (difference in rating scale is not considered)

$$sim(\vec{a}, \vec{b}) = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}| * |\vec{b}|}$$

- Certain similarity measures do consider the difference in the rating scale.

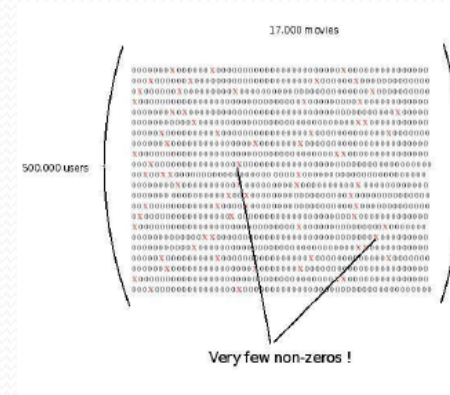
Amazon : Item – Item filtering [Implicit Rating]



This shows how the tracker collects the data from the users in to the matrix table. (The illustrated tracking method is a simplified version. You could also iterate the viewed items when a user view a new item to save all the item-to-item relation for the viewed items).

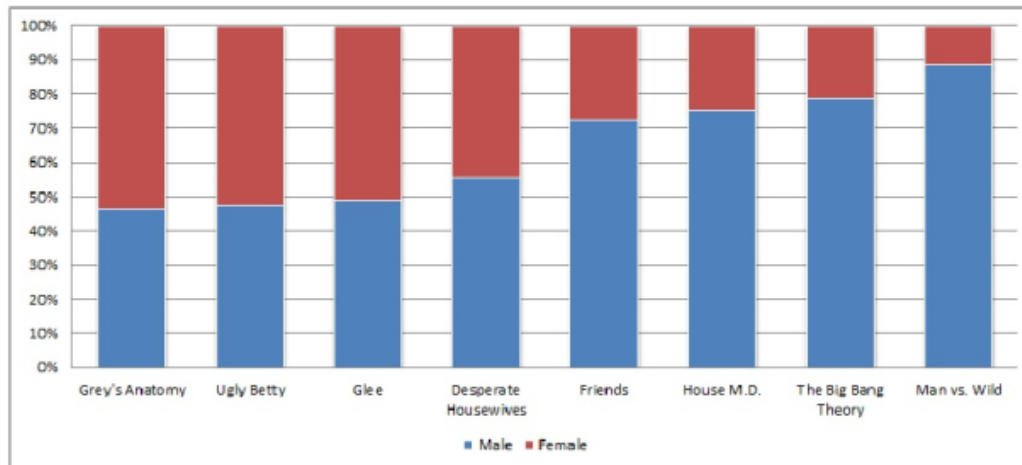
Challenges

- Data Sparsity
 - User/Item interactions are under 1%
 - Website usually contains Large collection of items, user ratings would be available for small percentage of them.
- Cold – start problem
 - How to recommend a new items?
 - Use content based information
 - What to recommend for new users?
 - Non personalized recommendation
 - Recommend highly rated items
 - Use profile of the user to recommend



Addressing user cold start

- Example: Gender and TV shows



Data comes from IMDB : <http://www.imdb.com/title/tt0412142/ratings>

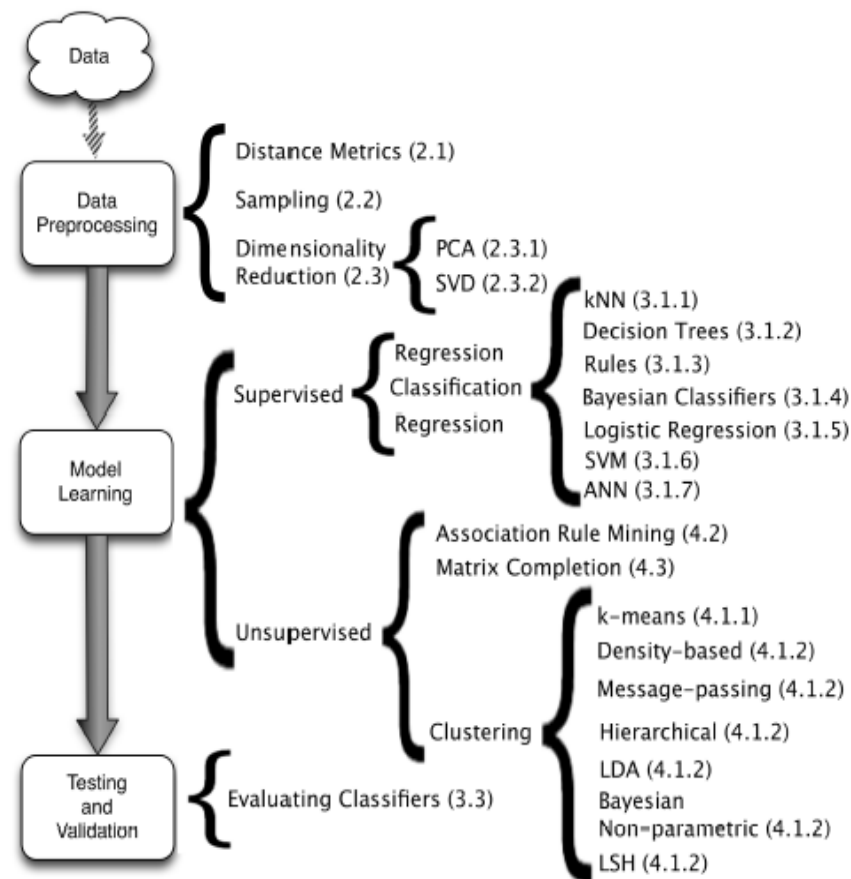
Contd..

- Data Scalability
 - Number of computations grows with increase in number of items and users
 - In order to address the issue, following preprocessing techniques are used by Amazon.
 - Calculate the item similarities in advance
 - Threshold : select only the items which are rated by at least n users
 - It limits the size of the neighborhood [but it affects the accuracy of the system]

Recommendation as Data Mining

The core of Recommendation engine can be assimilated as **general problem of data mining**.

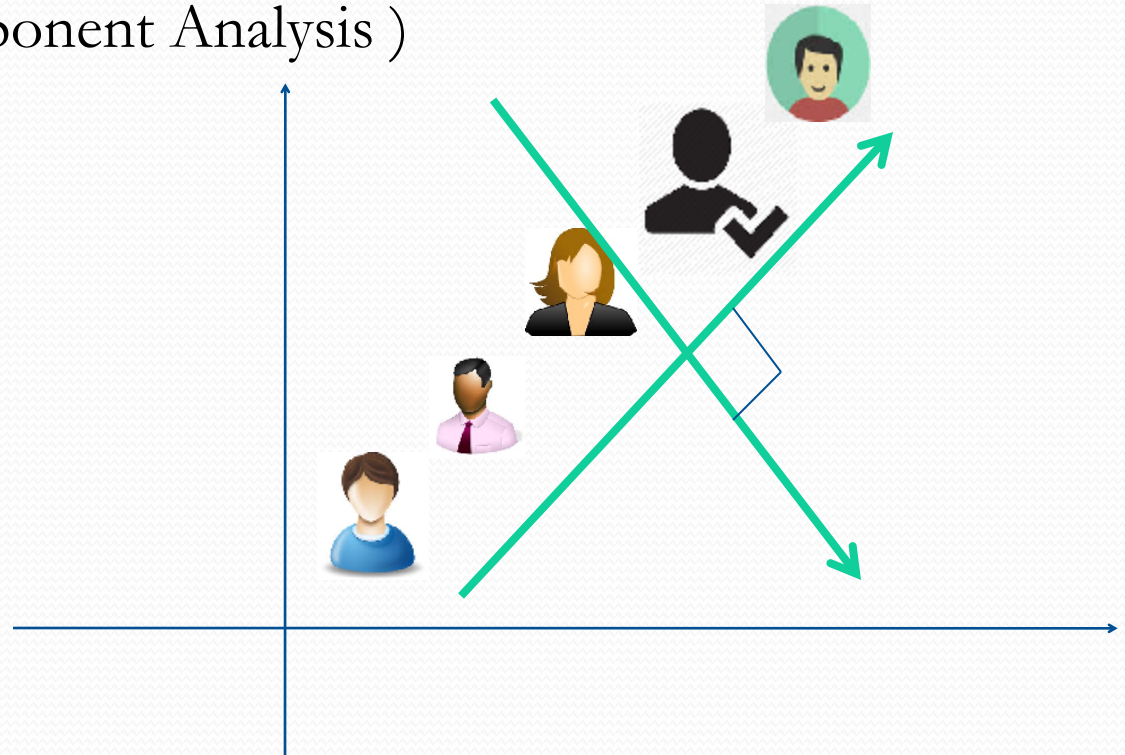
(Amatriain et al. Data Mining Methods for Recommender Systems in Recommender Systems Handbook)



Dimensionality Reduction

- PCA (Principal Component Analysis)

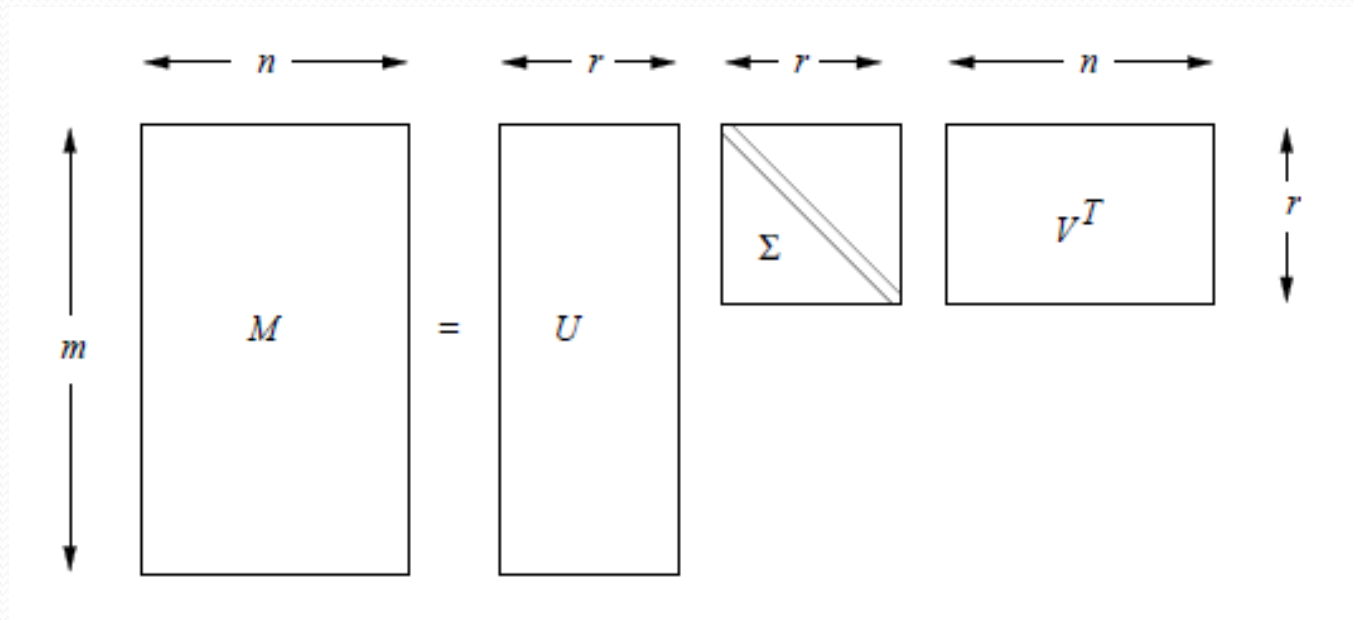
	I ₁	I ₂
	1	1
	2	2
	3	3
	4	4
	5	5



Contd..

- Instead of representing the data points in the normal axis.
- Represent the data points using the principal components.
- Direction which captures most variance in the data.
- How to identify the principal components?
 - Find the eigen values and eigen vector for the ratings matrix.
 - Eigen vectors – direction of the principal component.
 - Eigen values – variance associated with the specific direction
- Transform the original matrix to the transformed space where data points are represented with fewer number of principal components.

SVD (Singular Value Decomposition)



Interpreting SVD

	Matrix	Alien	Star Wars	Casablanca	Titanic
Joe	1	1	1	0	0
Jim	3	3	3	0	0
John	4	4	4	0	0
Jack	5	5	5	0	0
Jill	0	0	0	4	4
Jenny	0	0	0	5	5
Jane	0	0	0	2	2

Connects people to
concepts

Relates movies to
concepts

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 4 & 4 \\ 0 & 0 & 0 & 5 & 5 \\ 0 & 0 & 0 & 2 & 2 \end{bmatrix} = \begin{bmatrix} .14 & 0 \\ .42 & 0 \\ .56 & 0 \\ .70 & 0 \\ 0 & .60 \\ 0 & .75 \\ 0 & .30 \end{bmatrix} \begin{bmatrix} 12.4 & 0 \\ 0 & 9.5 \end{bmatrix} \begin{bmatrix} .58 & .58 & .58 & 0 & 0 \\ 0 & 0 & 0 & .71 & .71 \end{bmatrix}$$

M
 U
 Σ
 V^T

Provides strength of
each concept

Interpreting SVD

	Matrix	Alien	Star Wars	Casablanca	Titanic
Joe	1	1	1	0	0
Jim	3	3	3	0	0
John	4	4	4	0	0
Jack	5	5	5	0	0
Jill	0	2	0	4	4
Jenny	0	0	0	5	5
Jane	0	1	0	2	2

$$\begin{bmatrix} .13 & .02 & -.01 \\ .41 & .07 & -.03 \\ .55 & .09 & -.04 \\ .68 & .11 & -.05 \\ .15 & -.59 & .65 \\ .07 & -.73 & -.67 \\ .07 & -.29 & .32 \end{bmatrix} \begin{bmatrix} 12.4 & 0 & 0 \\ 0 & 9.5 & 0 \\ 0 & 0 & 1.3 \end{bmatrix} \begin{bmatrix} .56 & .59 & .56 & .09 & .09 \\ .12 & -.02 & .12 & -.69 & -.69 \\ .40 & -.80 & .40 & .09 & .09 \end{bmatrix}$$

U
 Σ
 V^T

Strength of the component is very small

Dropping out smallest Singular values

- Sum of the squares of the retained singular values should be at least 90% of the sum of squares of all singular values.

$$\begin{bmatrix} .13 & .02 \\ .41 & .07 \\ .55 & .09 \\ .68 & .11 \\ .15 & -.59 \\ .07 & -.73 \\ .07 & -.29 \end{bmatrix} \begin{bmatrix} 12.4 & 0 \\ 0 & 9.5 \end{bmatrix} \begin{bmatrix} .56 & .59 & .56 & .09 & .09 \\ .12 & -.02 & .12 & -.69 & -.69 \end{bmatrix} \\
 = \begin{bmatrix} 0.93 & 0.95 & 0.93 & .014 & .014 \\ 2.93 & 2.99 & 2.93 & .000 & .000 \\ 3.92 & 4.01 & 3.92 & .026 & .026 \\ 4.84 & 4.96 & 4.84 & .040 & .040 \\ 0.37 & 1.21 & 0.37 & 4.04 & 4.04 \\ 0.35 & 0.65 & 0.35 & 4.87 & 4.87 \\ 0.16 & 0.57 & 0.16 & 1.98 & 1.98 \end{bmatrix}$$

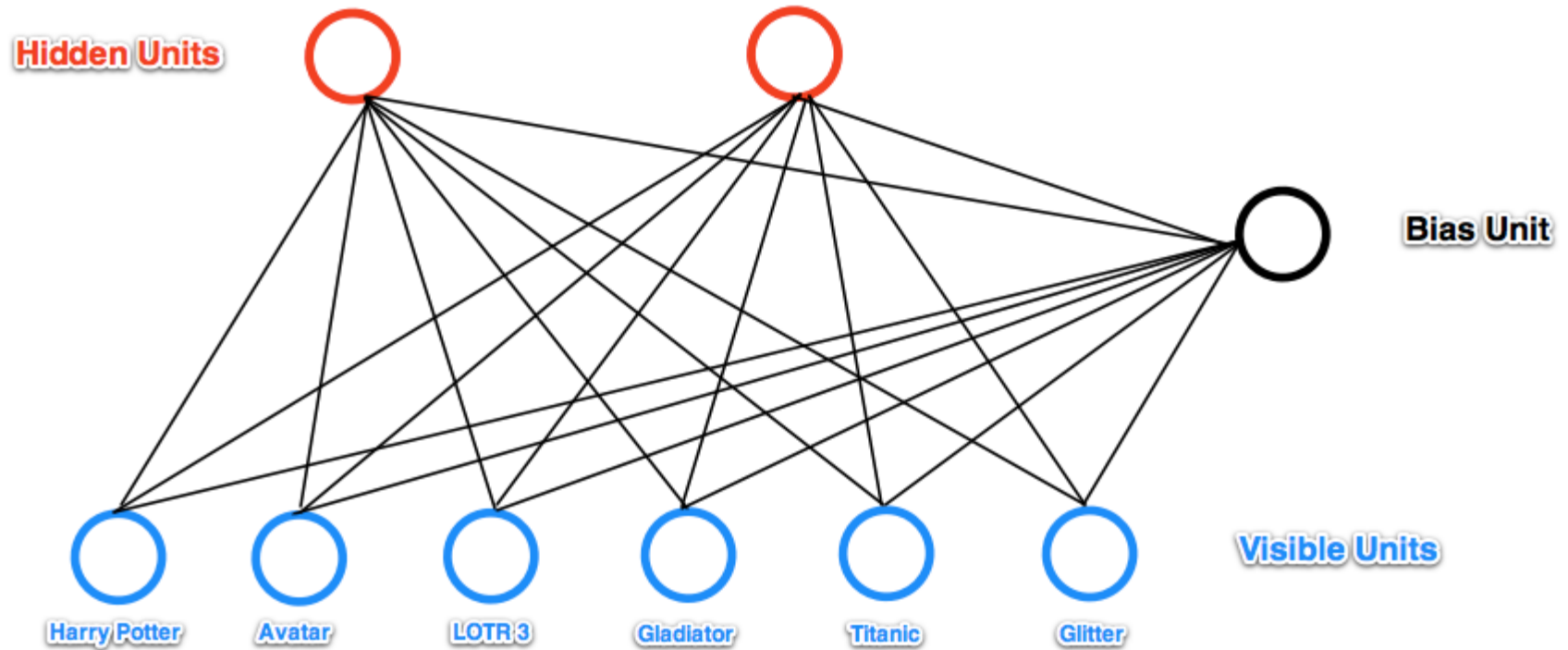
Prediction using SVD

- In order to find the items to be recommended to a new user
 $q=[4 \ 0 \ 0 \ 0 \ 0]$
- Map the vector to the concept space $x=qV$
 $X=[2.32 \ 0]$
- Map the result to the movie space by multiplying it with V^T
 $[1.35, 1.35, 1.35, 0, 0]$

Restricted Boltzmann Machine

- Binary version of factor analysis.
- Input is not a ratings matrix. Instead, the input is binary vector corresponding to whether the person liked the movie or not?
- RBM is a **stochastic neural network**
- It contains a layer of visible unit.
 - Number of nodes = number of movies
- The state of visible units depend on the user's movie preferences
- It also contains layer of hidden units which correspond to the concepts.

Contd..

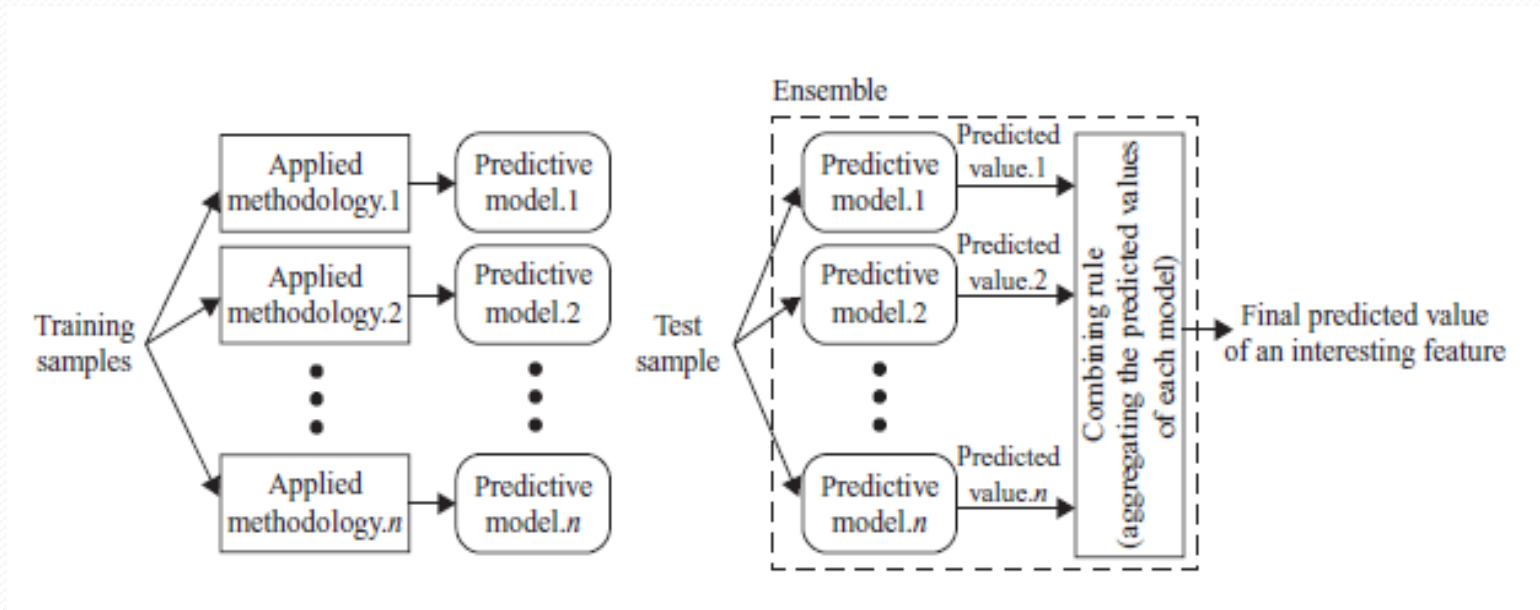


Training RBM

- On providing the movie preferences, visible units are activated.
- The activation state of the visible unit is fed as the input to the hidden units.
- Based on the input from visible units, activation state of hidden unit is determined.
- The resultant is fed as the input to the visible unit.
- The difference between the past and current activation states of the visible unit are used to adjust the weights and the threshold.

Ensemble models

- Making use of multiple models.
- Netflix uses ensemble of 107 algorithms



Open source tools-Collaborative Filtering

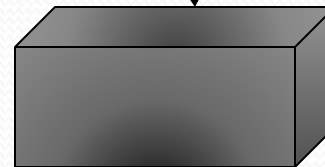
Software	Description	Language	URL
Apache Mahout	Hadoop ML library that includes Collaborative Filtering	Java	http://mahout.apache.org/
Cofi	Collaborative Filtering Library	Java	http://www.nongnu.org/cofi/
Crab	Components to create recommender systems	Python	https://github.com/muricoca/crab
easyrec	Recommender for web pages	Java	http://easyrec.org/
LensKit	Collaborative Filtering algorithms from GroupLens Research	Java	http://lenskit.grouplens.org/
MyMediaLite	Recommender system algorithms	C#/Mono	http://mloss.org/software/view/282/
SVDFeature	Toolkit for Feature based Matrix Factorization	C++	http://mloss.org/software/view/333/
Vogoo PHP LIB	Collaborative Filtering for personalized web sites	PHP	http://sourceforge.net/projects/vogoo/
recommenderlab	R library for developing and testing collaborative filtering systems	R	http://cran.r-project.org/web/packages/recommenderlab/index.html
Scikit-learn	Python module integrating classic ML algorithms in scientific Python packages (numpy , scipy , matplotlib)	Python	http://scikit-learn.org/stable/

Content based Recommendation

Find similar items based on the past preferences



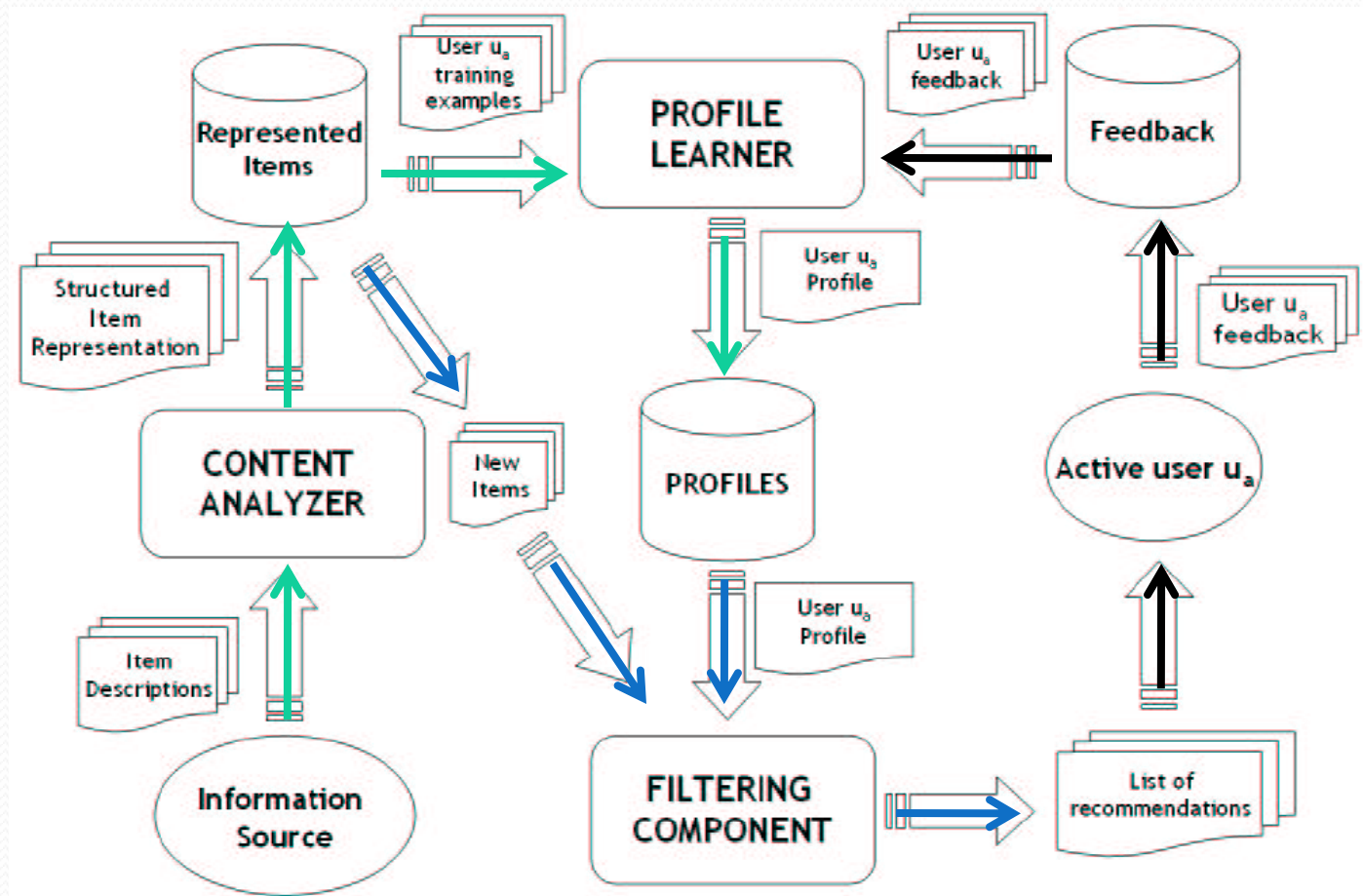
User Profile, Demographic/
Cognitive characteristics



Items	score
I ₁	0.87
I ₂	0.32
I ₃	0.1
I ₄	0.2

Item	Rate	Brand	Category

Content based Recommendation



Contd..

- Create profiles for both item and user.
- Mostly Content based recommendation systems are used to recommend the text documents ex. news articles or blogs.
- The item profile must capture the important characteristics of the document.
- Features : Structured and unstructured (text description)

Title	Genre	Author	Type	Price	Keywords
The Night of the Gun	Memoir	David Carr	Paperback	29.90	Press and journalism, drug addiction, personal memoirs, New York
The Lace Reader	Fiction, Mystery	Brunonia Barry	Hardcover	49.90	American contemporary fiction, detective, historical
Into the Fire	Romance, Suspense	Suzanne Brockmann	Hardcover	45.90	American fiction, murder, neo-Nazism

Contd..

- Unstructured – Identification of words for characterize the topic of a given document.
 - Eliminate stop words.
 - Compute TF-IDF score for the remaining words.
 - The words with highest TF-IDF score is used to characterize the document.
- The user profile is created using the ratings provided by the user for the documents
- Similarity measure: cosine similarity is used to find the relevance of item profile and user profile.

Recommending items – (Query based Retrieval)

- Users are allowed to provide feedback on relevance of the document.
- Queries are automatically extended with additional weight to the relevant documents.

Probabilistic Models

- The problem of recommendation is modeled as a two class problem.
- Model:
 - 2 classes: Like/Dislike
 - Simple Boolean representation
 - Calculate the probability that the item is liked / disliked by the user based on Bayes theorem.

Doc-ID	recommender	intelligent	learning	school	Label
1	1	1	1	0	1
2	0	0	1	1	0
3	1	1	0	0	1
4	1	0	1	1	1
5	0	0	0	1	0
6	1	1	0	0	?

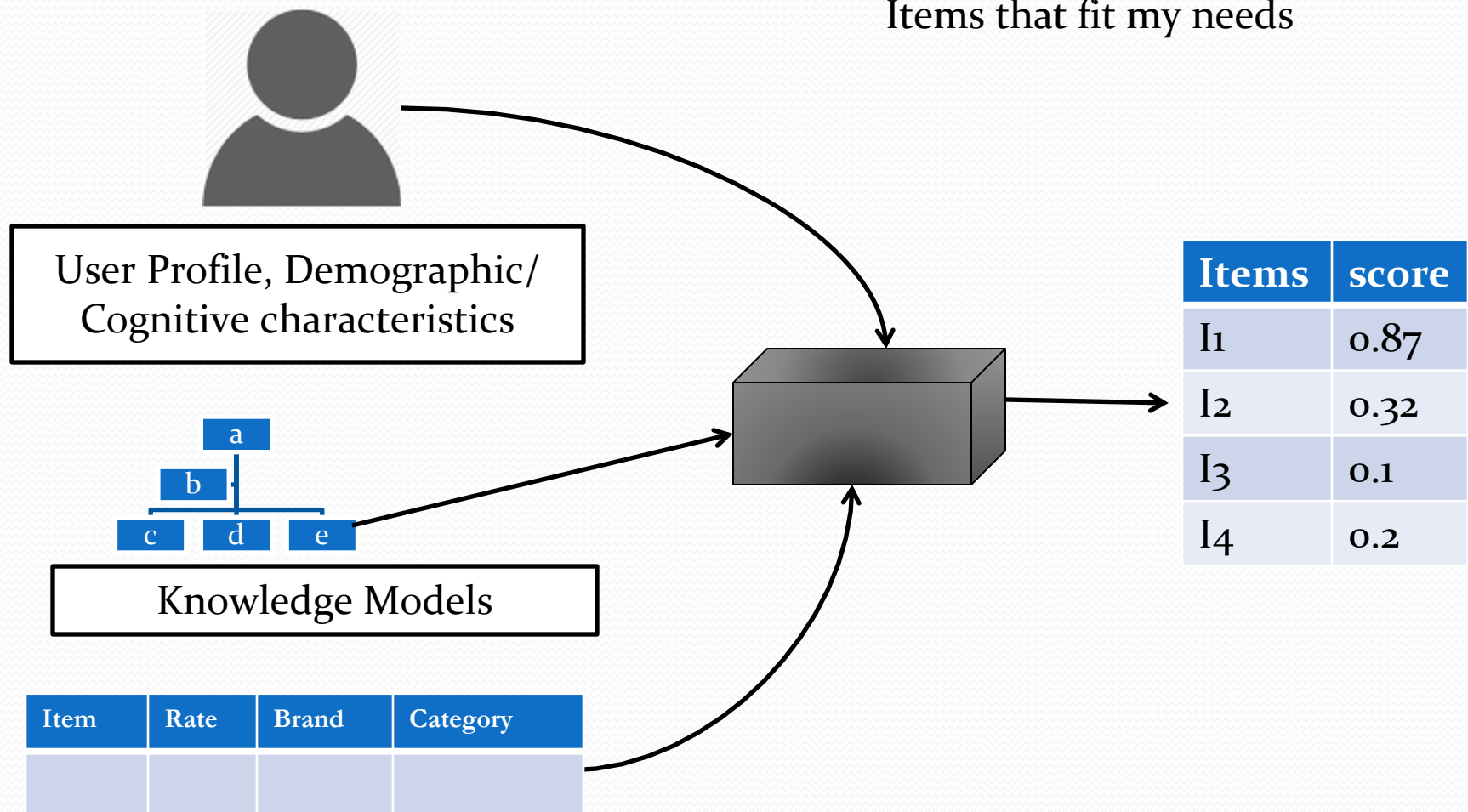
$$\begin{aligned}P(X|\text{Label}=1) &= P(\text{recommender}=1|\text{Label}=1) \times \\&\quad P(\text{intelligent}=1|\text{Label}=1) \times \\&\quad P(\text{learning}=0|\text{Label}=1) \times P(\text{school}=0|\text{Label}=1) \\&= 3/3 \times 2/3 \times 1/3 \times 2/3 \\&\approx 0.149\end{aligned}$$

Limitations

- Overspecialization
- Limited Content Analysis
- Keywords may not be sufficient.

Knowledge based Recommendation


Items that fit my needs



Knowledge Based Recommendation

- Constraint Based
 - Explicitly defined conditions.
- Case Based
 - Similarity to specified requirements

Constraint Based (WeeVis)



- Main page
- Recent changes
- Random page
- Help

▼ Tools

- What links here
- Related changes
- Special pages
- Printable version
- Permanent link
- Page information

Log in Request account

Page DiscussionRead View source View historySearch

SmartphoneRecommender

KU Wissensverarbeitung(Expertensysteme) SS2014

Questions

Which operating system do you prefer?
no answer

Which case type do you prefer?
no answer

Does your Smartphone need to have a fingerprint sensor?
no answer

Which display size do you prefer?
no answer

How many RAM you need?
no answer

Solutions	Support
iPhone 5	100%
iPhone 5c	100%
iPhone 5s	100%
Nokia Lumia 630	100%
Samsung Galaxy S5	100%
Sony Xperia Z1	100%

Limitations

- Cost of Knowledge acquisition
- Accuracy of models

Research Challenges

- Scalability
- Proactive recommender systems
- Optimization of recommendations
- Distributed recommender system

Datasets

- <https://gist.github.com/entaroadun/1653794>



Thank you