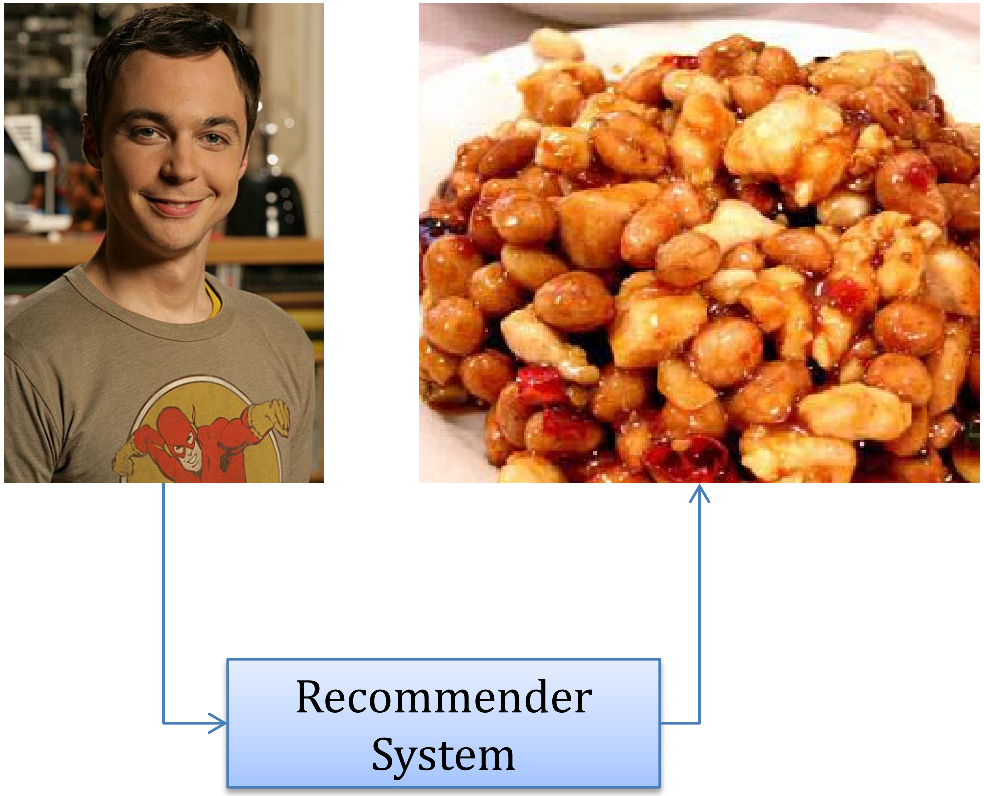
# Recommender System: Algorithms & Architecture

xiangliang@hulu.com

# Outline

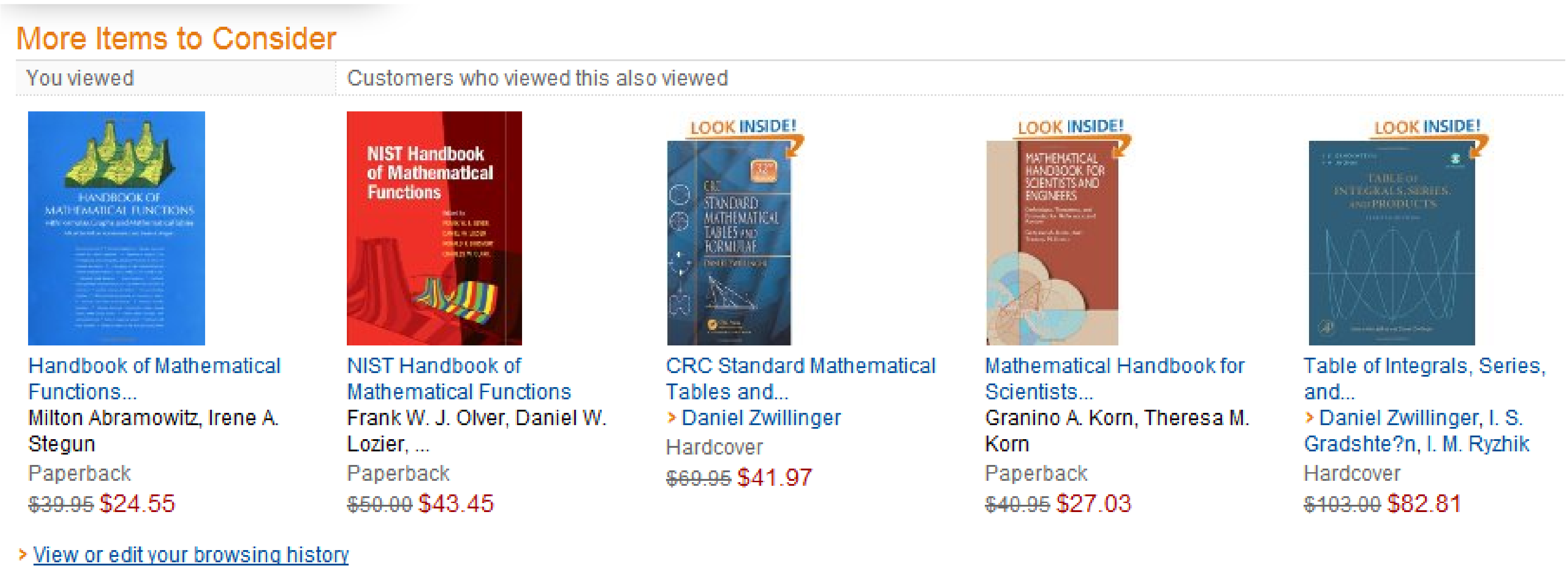
- Problem
- Data
- Algorithms
- Cold start
- Architecture



Recommender System

# Problem

***Recommend items to users to make user, content partner, websites happy!***



More Items to Consider

You viewed

Customers who viewed this also viewed

Handbook of Mathematical Functions...
Milton Abramowitz, Irene A. Stegun
Paperback
~~$39.95~~ $24.55

NIST Handbook of Mathematical Functions
Frank W. J. Olver, Daniel W. Lozier, ...
Paperback
~~$50.00~~ $43.45

CRC Standard Mathematical Tables and...
› Daniel Zwillinger
Hardcover
~~$69.95~~ $41.97

Mathematical Handbook for Scientists...
Granino A. Korn, Theresa M. Korn
Paperback
~~$40.95~~ $27.03

Table of Integrals, Series, and...
› Daniel Zwillinger, I. S. Gradshte?n, I. M. Ryzhik
Hardcover
~~$103.00~~ $82.81

› View or edit your browsing history

# Data

- User behaviors data

| Behavior | User | Size |
|---|---|---|
| Page view | All user | Very Large |
| Watch video | All user | Large |
| Favorite | Register user | Middle |
| Vote | Register user | Middle |
| Add to playlist | Register user | Small |
| Facebook like | Register user | Small |
| Share | Register user | Small |
| Review | Register user | Small |

# Data

- Which data is most important
  - Main behavior in the website
  - All user can have such behavior
  - Cost
  - Reflect user interests on items

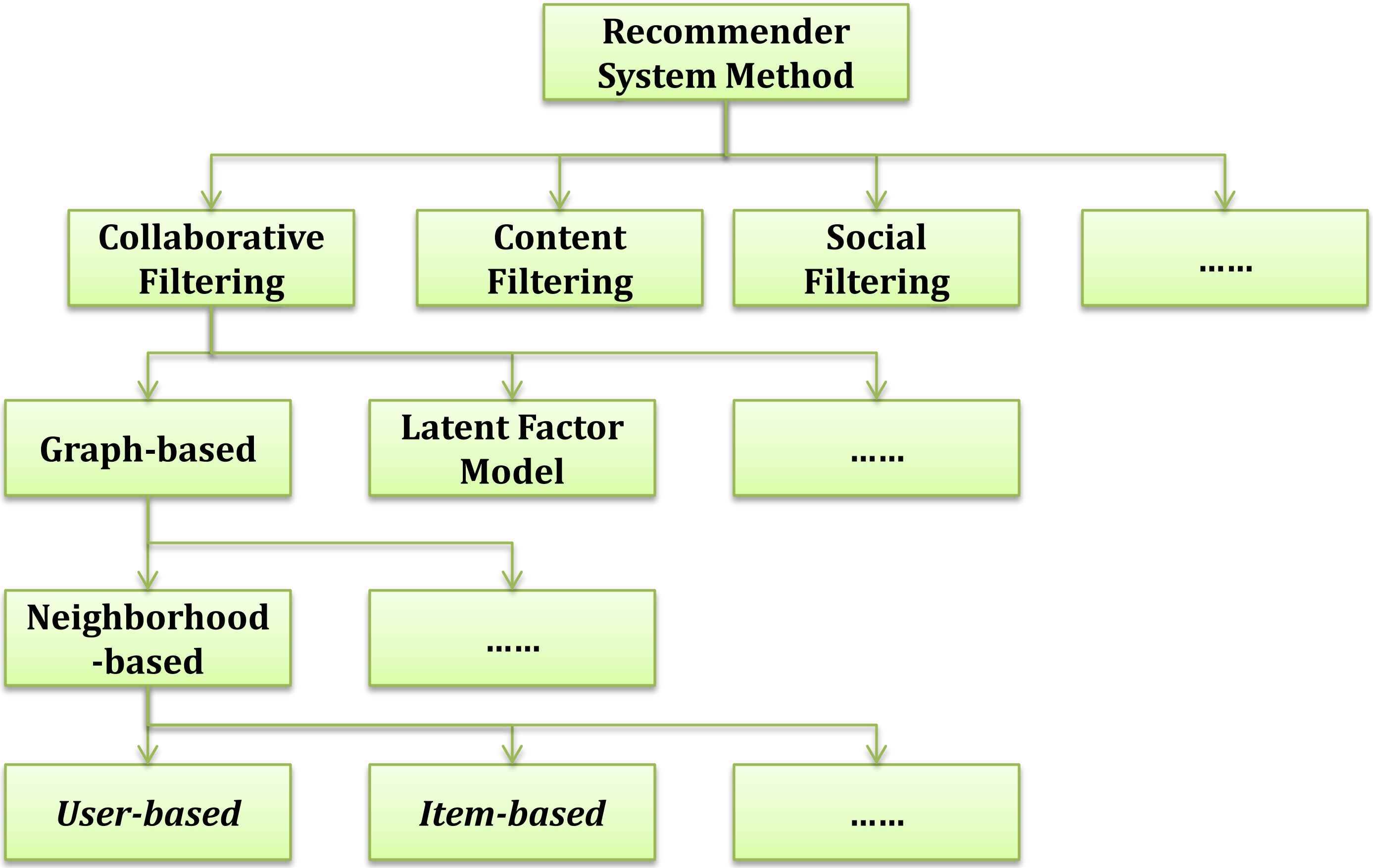
| Behavior | User | Size |
|---|---|---|
| Page view | All user | Very Large |
| Watch video | All user | Large |
| Favorite | Register user | Middle |
| Vote | Register user | Middle |
| Add to playlist | Register user | Small |
| Facebook like | Register user | Small |
| Share | Register user | Small |
| Review | Register user | Small |

# Data

- Data Structure
  - User ID
  - Item ID
  - Behavior Type
  - Behavior Content
  - Context
    - Timestamp
    - Location
    - Mood

**Sheldon watch Star Trek with his friends at home**

# Algorithms



Recommender System Method

Collaborative Filtering

Content Filtering

Social Filtering

……

Graph-based

Latent Factor Model

……

Neighborhood -based

……

*User-based*

*Item-based*

……

# Neighborhood-based

- User-based
  - Digg
- Item-based
  - Amazon, Netflix, YouTube, Hulu, …

# User-based

- Algorithm  u → u
  - For user u, find a set of users S(u) have similar preference as u.
  - Recommend popular items among users in S(u) to user u.

# User-based CF

$$p_{ui} = \sum_{v \in S(u,K) \cap N(i)} w_{uv} r_{vi}$$

$$\mathrm{w}_{uv} = \frac{|N(u) \cap N(v)|}{|N(u) \cup N(v)|}$$

# Item-based

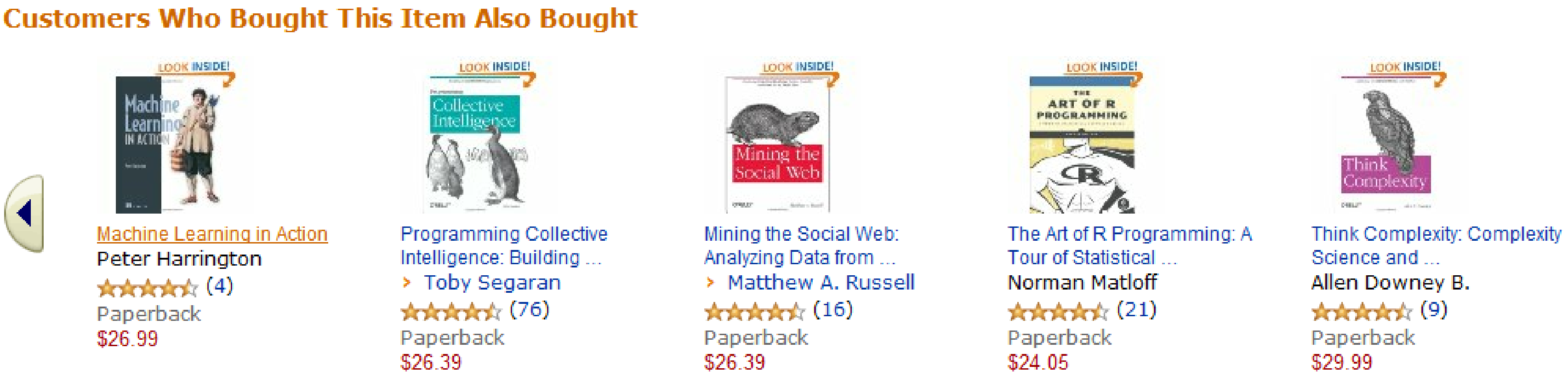
- Algorithm
  - For user u, get items set N(u) this user like before.
  - Recommend items which are similar to many items in N(u) to user u.

# Item-based CF

$$p_{ui} = \sum_{j \in S(i,K) \cap N(u)} w_{ji} r_{uj}$$

$$\mathrm{w}_{ij} = \frac{\left| N(i) \cap N(j) \right|}{\left| N(i) \cup N(j) \right|}$$

# Item-based CF



Customers Who Bought This Item Also Bought

Machine Learning in Action
Peter Harrington
(4)
Paperback
$26.99

Programming Collective Intelligence: Building ...
› Toby Segaran
(76)
Paperback
$26.39

Mining the Social Web: Analyzing Data from ...
› Matthew A. Russell
(16)
Paperback
$26.39

The Art of R Programming: A Tour of Statistical ...
Norman Matloff
(21)
Paperback
$24.05

Think Complexity: Complexity Science and ...
Allen Downey B.
(9)
Paperback
$29.99

Why not use $$w_{ij} = \frac{|N(i) \cap N(j)|}{|N(i)|}$$ ?

# Neighborhood-based

- User-based vs. Item-based

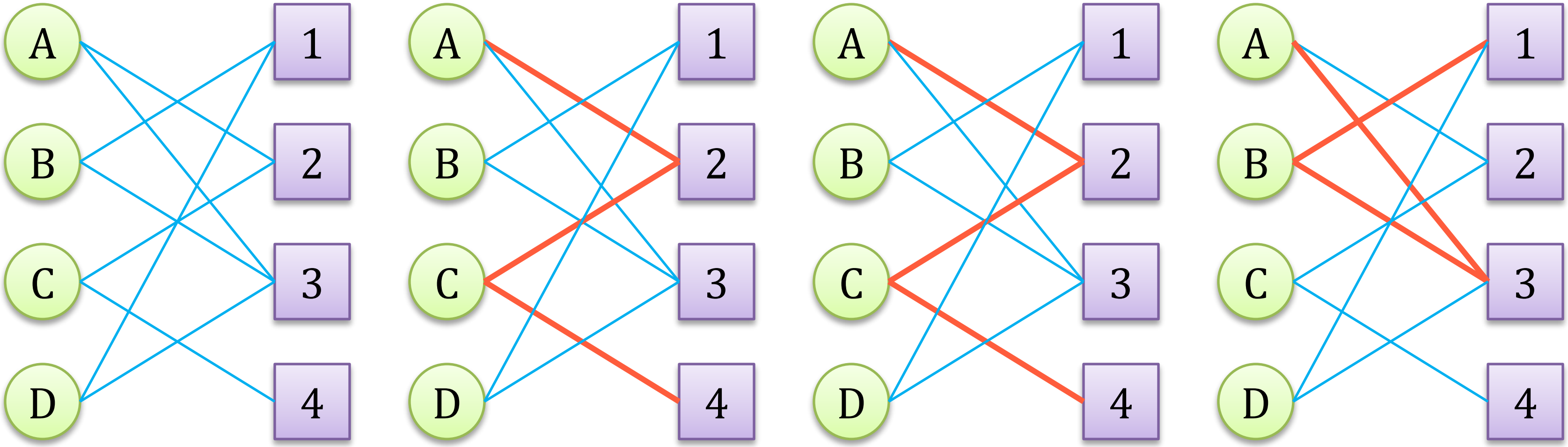
| | User-based | Item-based |
|---|---|---|
| Scalability | Bad when user size is large<br>user | Bad when item size is large<br>Item size sensitive. |
| Explanation | Bad | Good |
| Novelty | Bad | Good |
| Coverage | Bad | Good |
| Cold start | Bad for new users | Bad for new items |
| Performance | Need to get many users history | Only need to get current user's history |

# References

- Amazon.com Recommendations item-to-item Collaborative Filtering.
- Empirical Analysis of Predictive Algorithms for Collaborative Filtering.

# Graph-based

- Users' behaviors on items can be represented by bi-part graph.

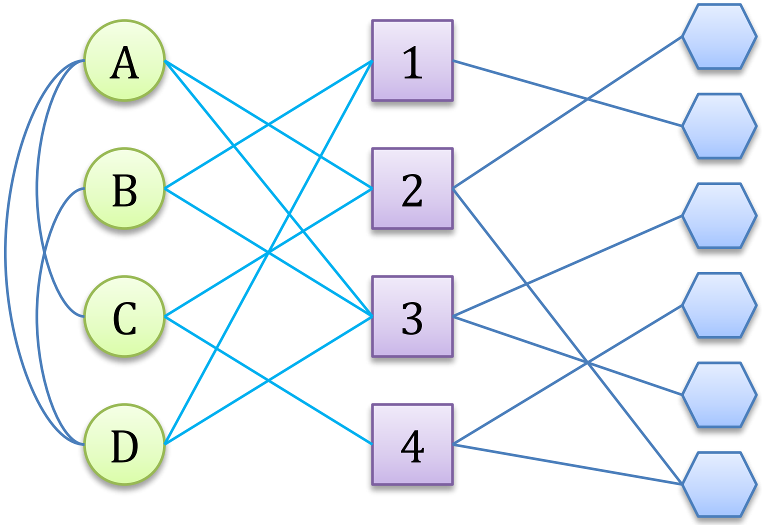


# Graph-based

- Two nodes will have high relevance if
  - There are many paths in graph between two nodes.
  - Most of paths between two nodes is short.
  - Most paths do not go through nodes with high out-degree.

# Graph-based

- Advantage
  - Heterogeneous data
    - Multiple user behaviors
    - Social Network
    - Context (Time, Location)
- Disadvantage
  - Statistical-based
  - High cost for long path

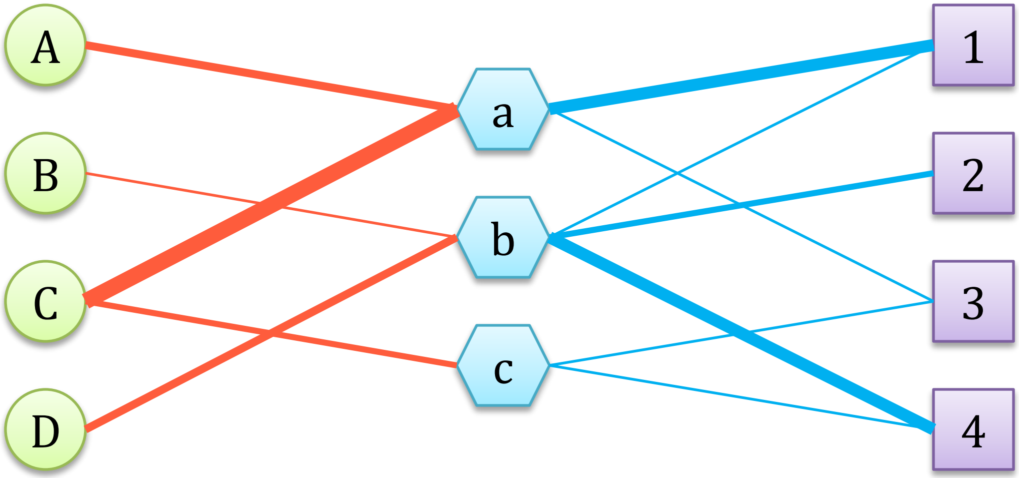


# References

- A Graph-based Recommender System for Digital Library.
- Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation.

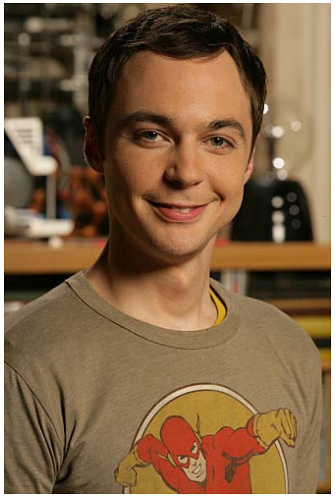
# Latent Factor Model

- Users and items are connect by latent features.



# Latent Factor Model

$$\hat{r}_{ui} = \sum_{k} p_{uk} q_{ik}$$



| | |
|---|---|
| Science Fiction | 0.5 |
| Universe | 0.9 |
| Physical | 0.8 |
| Space Travel | 0.8 |
| Animation | 0.3 |
| Romance | 0.0 |

| | |
|---|---|
| Science Fiction | 0.9 |
| Universe | 0.9 |
| Physical | 0.5 |
| Space Travel | 0.7 |
| Animation | 0.1 |
| Romance | 0.0 |

# Latent Factor Model

- How to get p, q?

$$\min \sum_{(u,i)} (r_{ui} - \sum_{k} p_{uk} q_{ik})^2 + \lambda(\|p_{\mathrm{u}}\|^2 + \|q_i\|^2)$$

$$p_{uk} += \alpha(e_{ui} q_{ik} - \lambda p_{uk})$$

$$q_{ik} += \alpha(e_{ui} p_{uk} - \lambda q_{ik})$$

# Latent Factor Model

- How to define $r_{ui}$
  - Rating prediction
  - Top-N recommendation
    - Implicit feedback data: only have positive samples and missing values, how to select negative samples?

# Latent Factor Model

| 1 (Sci-fi) | 2 (Crime) | 3 (Family) | 4 (Horror) |
|---|---|---|---|
| The invisible Man | Jaws | 101 Dalmatians | The Blair Witch Project |
| Frankenstein Meets the Wolf Man | Lethal Weapon | Back to the Future | Pacific Heights |
| Godzilla | Total Recall | Groundhog Day | Stir of Echoes |
| Star Wars VI | Reservoir Dogs | Tarzan | Dead Calm |
| The Terminator | Donnie Brasco | The Aristocats | Phantasm |
| Alien | The Fugitive | The Jungle Book 2 | Sleepy Hollow |
| Alien 2 | La shou Shen tan | Antz | The Faculty |

# Latent Factor Model

- Advantage
  - High accuracy in rating prediction
  - Auto group items
  - Scalability is good
  - Learning-based
- Disadvantage
  - Incremental updating
  - Real-time
  - Explanation

# References

- http://www.informatik.uni-trier.de/~ley/db/indices/a-tree/k/Koren:Yehuda.html

# Cold Start

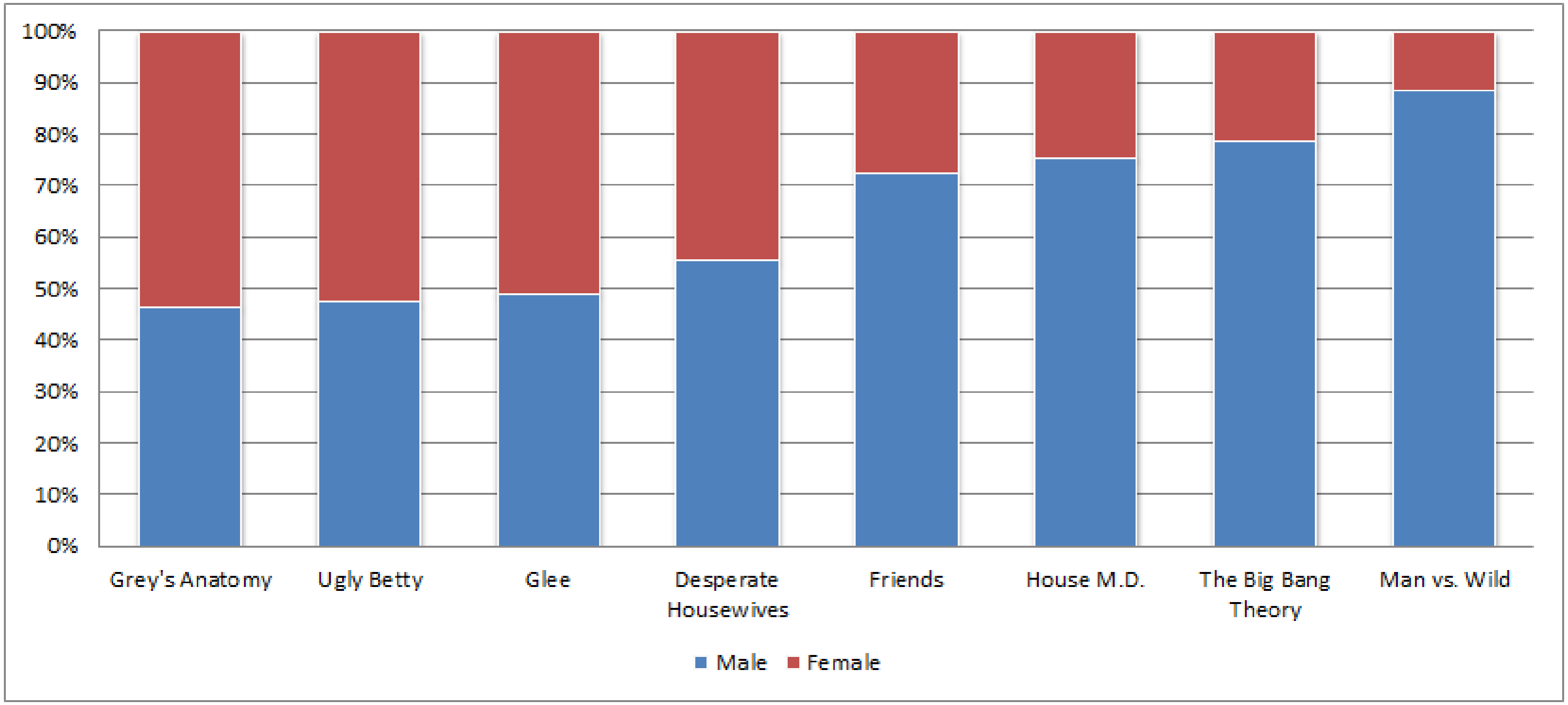
- Problems
  - User cold start : new users
  - Item cold start : new items
  - System cold start : new systems

# User Cold Start

- How to recommend items to new users?
  - Non-personalization recommendation
    - Most popular items
    - Highly Rated items
  - Using user register profile (Age, Gender, …)

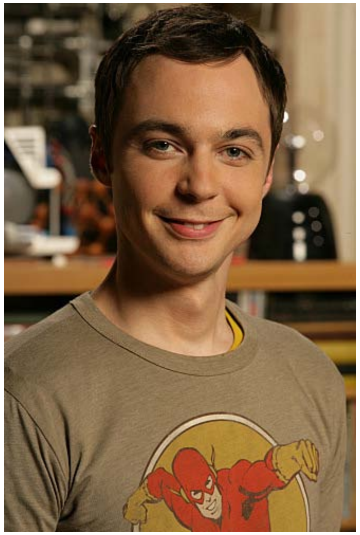
# User Cold Start

- Example: Gender and TV shows



Data comes from IMDB : [http://www.imdb.com/title/tt0412142/ratings](http://www.imdb.com/title/tt0412142/ratings)

# User Cold Start



Male
Age : 20-30
Theoretical physicist
Doctor
American
Irreligious

# How to get user interest quickly

- When new user comes, his feedback on what items can help us better understand his interest?
  - Not very popular
  - Can represent a group of items
  - Users who like this item have different preference with users who dislike this item

# Item Cold Start

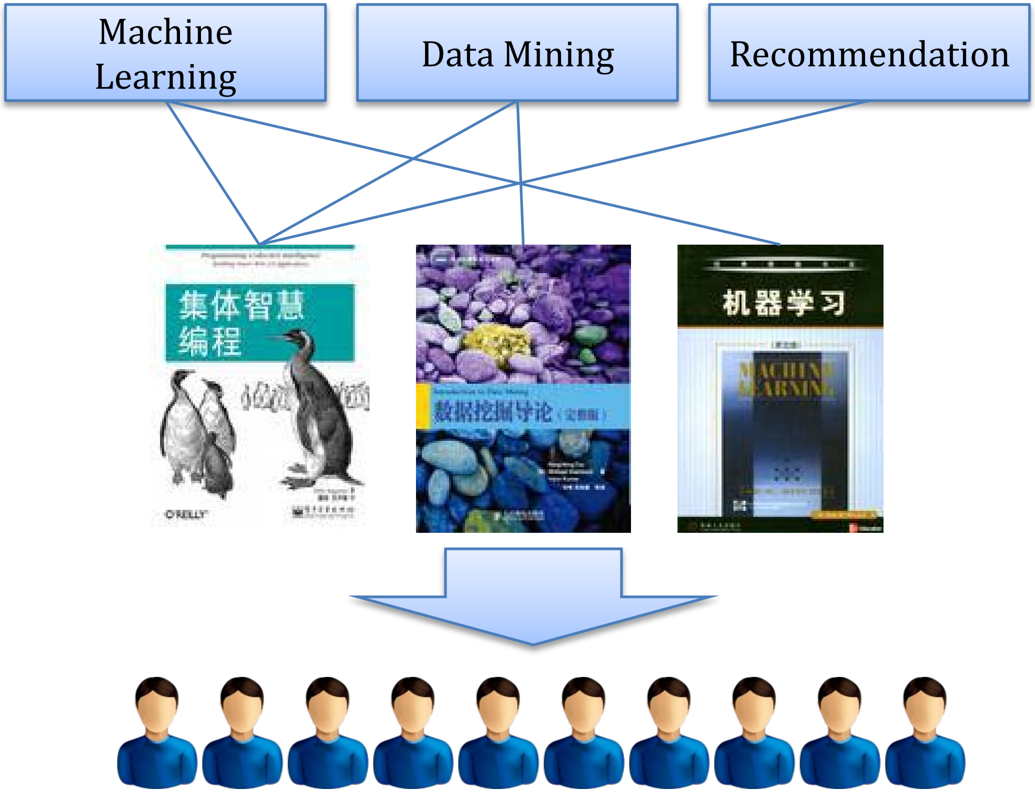
- How to recommend new items to user?
  - Do not recommend

新书速递 - 虚构类 · · · · · · （查看更多）

**How to recommend news??**

# Item Cold Start

- How to recommend new items to user?
  - Using content information


Machine Learning

Data Mining

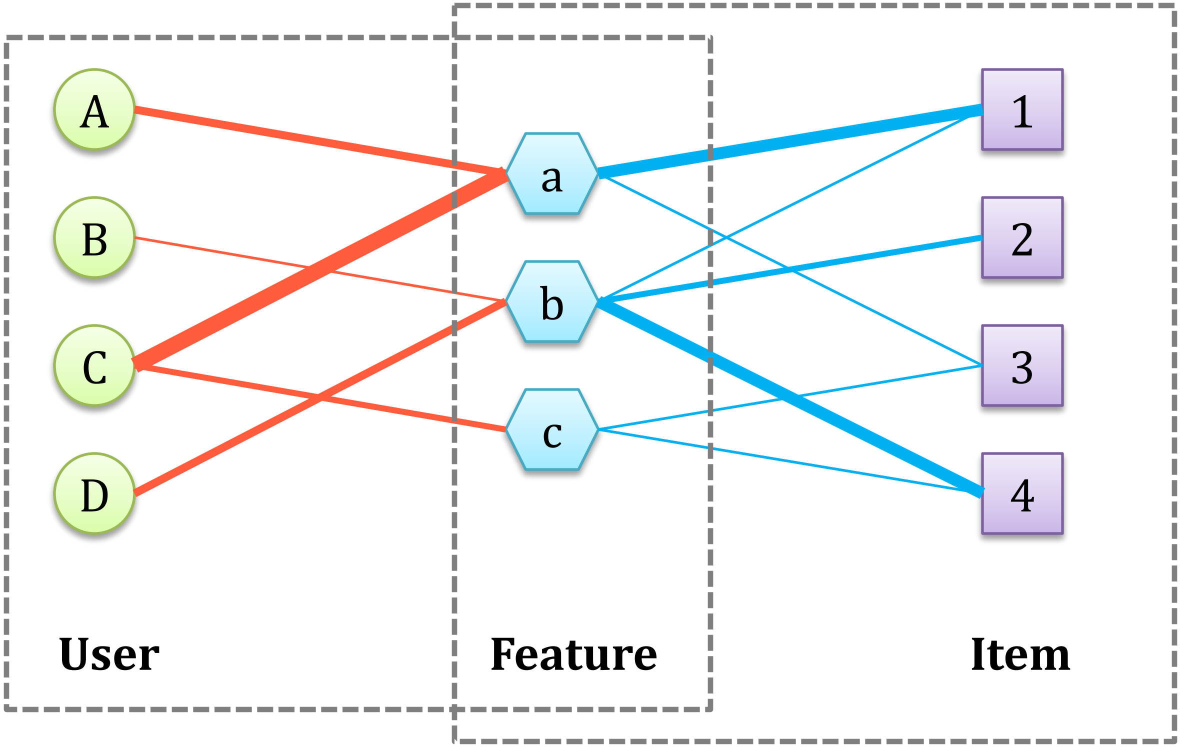
Recommendation

# System Cold Start

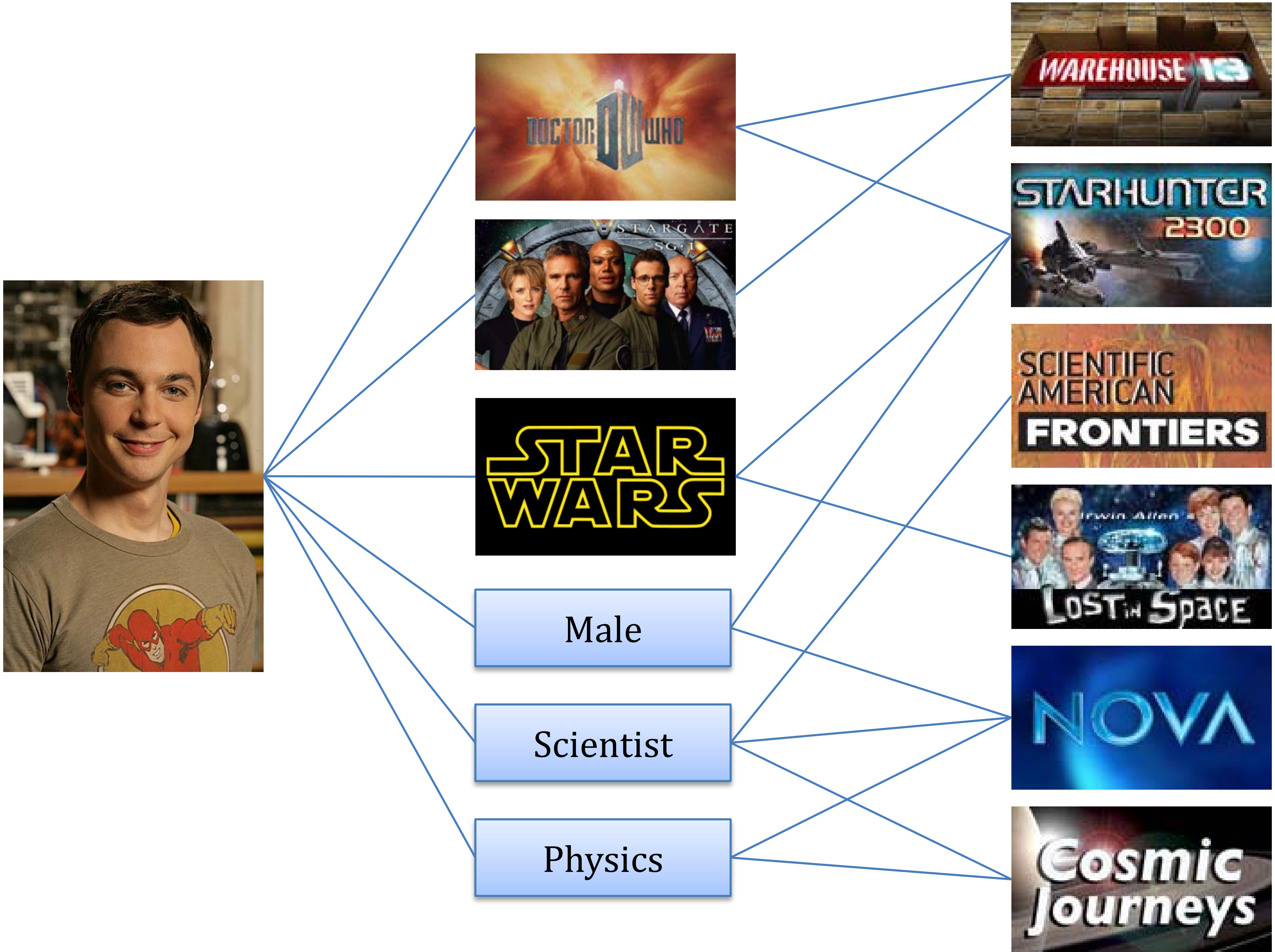
- How to design recommender system when there is no user?
  - Pandora : Music Genome Project
  - Jinni : Movie Genome Project

# Architecture

- Feature-based recommendation framework:



# Architecture



Male

Scientist

Physics

# Architecture

- Advantage:
  - Heterogeneous data
  - Reasonable Explanation
- Disadvantage:
  - Do not support user-based methods

# Open Questions

- How to weight multiple behaviors?
- How to improve diversity, novelty?
- How to build feedback loop?

?

RL

Thanks!