Recommendation Systems

◆Predict items a user may be interested in based on information about the user and the items
◆An effective way to help people cope with information overload
◆Examples: Amazon, Netflix, Tivo, ...

Problem Definitions

◆Prediction
  ◆ Given: a user and \( k \) items
  ◆ Return: predicted rating for each item

◆Recommendation
  ◆ Given: a user
  ◆ Return: \( k \) items from the database with the highest predicted rating

Categories of Filtering Techniques

◆Cognitive (content-based) filtering
◆Economic filtering
◆Collaborative (social) filtering
  ◆ Rate items based on the evaluation of other users

Collaborative vs. Cognitive

◆Support for filtering items whose content is not easily analyzed by automated process
◆The ability to filter items based on quality and taste
◆The ability to provide serendipitous recommendations

CF – Basic Assumptions

◆Items are evaluated by users explicitly or implicitly
  ◆ Ratings, reviews
  ◆ Purchases, browsing behaviors
  ◆ ...
◆We may map explicit and implicit evaluations to a rating scale, e.g. 1-5.

So shall we simply recommend the items with the highest average rating??
Problem Formulation

User-Item Matrix

<table>
<thead>
<tr>
<th>Item</th>
<th>Ken</th>
<th>Lee</th>
<th>Meg</th>
<th>Nan</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>2</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>5</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>??</td>
<td>2</td>
<td>5</td>
<td></td>
</tr>
</tbody>
</table>

So what would be Ken's rating for Item 6??

Predicted Rating

\[ p_{x,i} = \bar{r}_i + \sum_{u} \left( r_{u,i} - \bar{r}_u \right) \times w_{x,u} \]

So what is \( p_{ken,6} \)??

Algorithm Quality Metrics

Coverage – percentage of items for which the system can produce a prediction

Accuracy
  - Statistical metrics
    - Mean Absolute Error (MAE)
  - Decision-support metrics

Efficiency
  - Throughput – number of recommendations per second

Variants and Optimizations

Similarity measure
Significance weighting
Item rating variance
Neighborhood selection
Rating normalization
Similarity Measures
- Pearson Correlation
- Spearman Correlation
- Cosine similarity
- Entropy
- Mean-squared-difference
- ...

Significance Weighting
- Weight users in addition to the similarity measure

Item Rating Variance
- Some items are more telling about tastes than others

Neighborhood Selection
- Select a subset of the users for better performance and accuracy.

Rating Normalization
- No normalization
- Deviation from mean
- Z-score

Other Recommendation Algorithms
- Combine collaborative and content-based filtering
- Item-item collaborative filtering
- Bayesian networks
CF Libraries

- Taste – http://taste.sourceforge.net/
- COFE – http://eecs.oregonstate.edu/iis/CoFE/

Non-personalized Recommendation

- What if the user is new to the site?
- What if the site itself is new, i.e. no previous user transactions?

Sales Transactions

| 11: Beef, Chicken, Milk |
| 12: Beef, Cheese        |
| 13: Cheese, Boots       |
| 14: Beef, Chicken, Cheese |
| 15: Beef, Chicken, Clothes, Cheese, Milk |
| 16: Chicken, Clothes, Milk |
| 17: Chicken, Milk, Clothes |

Amazon-like recommendation:
Users who purchased milk also purchased the following items:
- Clothes
- Chicken

Association Rule Mining

- \( \{i_1, i_2, ..., i_n\} \rightarrow j \)
- Confidence: the probability of finding item \( j \) in a transaction that has \( \{i_1, i_2, ..., i_n\} \)
- Support: the number of transactions that have \( \{i_1, i_2, ..., i_n\} \) and \( j \)

A-Priori Algorithm

- Observation: A set of items \( x \) has support \( s \), then each subset of \( x \) must have support at least \( s \).
- Example: find the association rules that have at least 20% support and 50% confidence

Item Similarity under Vector-Space Model

- Each unique term is a dimension
- Each document is a vector
- Similarity
  - Euclidean distance
  - Cosine similarity measure
References