INTRODUCTION TO RECOMMENDER SYSTEMS AND THEIR EVALUATION

Olga Popova
AGENDA

Definition
Factors defining recommender systems
Introduction
Properties of recommender systems
Evaluating algorithms
Offline experiment
User study experiment
Online experiment
Metrics to measure prediction accuracy and their applications
Conclusion
Bibliography
AGENDA

**Definition**

Introduction
Factors defining recommender systems
Properties of recommender systems
Evaluating algorithms
Offline experiment
User study experiment
Online experiment
Metrics to measure prediction accuracy and their applications
Conclusion
Bibliography
There is an extensive class of Web applications that involve predicting user responses to options. Such a facility is called a recommendation system [1]
AGENDA

Definition
Factors defining recommender systems
Introduction
Properties of recommender systems
Evaluating algorithms
Offline experiment
User study experiment
Online experiment
Metrics to measure prediction accuracy and their applications
Conclusion
Bibliography
INTRODUCTION - FACTORS DEFINING RECOMMENDER SYSTEMS

[2] Deepak Agarwal & Bee-Chung Chen @ ICML'11
Recommender problems & [3] Gediminas Adomavicius,
Alexander Tuzhilin, Context-Aware Recommender Systems
INTRODUCTION

Initial goal: predict/recommend precisely

Several other needs emerged during the last time:
1. discover new
2. preserve their privacy
3. fast-responding system
4. ...

Properties can be measured with help of special metrics
AGENDA

Definition
Factors defining recommender systems
Introduction
Properties of recommender systems
Evaluating algorithms
Offline experiment
User study experiment
Online experiment
Metrics to measure prediction accuracy and their applications
Conclusion
Bibliography
PROPERTIES OF RECOMMENDATION SYSTEMS

- Prediction accuracy
AGENDA

Definition
Factors defining recommender systems
Introduction
Properties of recommender systems
Evaluating algorithms
Offline experiment
User study experiment
Online experiment
Metrics to measure prediction accuracy and their applications
Conclusion
Bibliography
EVALUATING ALGORITHMS

Evaluation in terms of properties:
- define proper qualities and try to enhance on them
  - Show that the property is relevant
  - Design an algorithm to improve on these properties

Trade-off between a set of properties:
- one gets better, another gets worse

Fix all the other variables that are not tested
- (using the same dataset)

Generalization
- (what’s good for one application/dataset, is not for another one)

Data biases
- Exclude items with low counts/try to eliminate biases

[Diagram showing Offline, User study, Online]
AGENDA

- Definition
- Factors defining recommender systems
- Introduction
- Properties of recommender systems
- Evaluating algorithms
  - Offline experiment
  - User study experiment
  - Online experiment
- Metrics to measure prediction accuracy and their applications
- Conclusion
- Bibliography
Offline experiment is done on the base of pre-collected data of users rating or choosing items. Data can be used for simulation of users behavior.

Positive:
- cheap
- helps to filter out inappropriate approaches

Negative:
- answers narrow set of questions (example: prediction power)
USER STUDY

- Recruit test persons and ask them to implement some tasks

**positive**
- Collects quantitative data
- Collects qualitative data
- Influence of...

**to consider**
- False questioning can influence user’s opinion
- Hypothesis should be...
ONLINE EXPERIMENT

System used by real users performing real tasks with lots of data

positive
• Real user behavior
• Possible to influence user
• User compares

to consider
• Better to fix the interface to compare algorithms without biases
• Costly
ONLINE VS. OFFLINE

Higher offline evaluation accuracy >>> better perceived quality by users

Not really

Analyzing only towards offline evaluation accuracy brings the risk of tuning an algorithm to recommend items which could potentially not be liked by the user in a real-world scenario. [5]
AGENDA

Definition
Factors defining recommender systems
Introduction
Properties of recommender systems
Evaluating algorithms
Offline experiment
User study experiment
Online experiment
Metrics to measure prediction accuracy and their applications
Conclusion
Bibliography
SCENARIOS

Platforms

Use cases

1) Comparison: real score to the predicted one
2) Evaluation of created lists of recommendations (films)
3) Evaluation of recommendations order
4) Definition of users with close tastes

1) Evaluation on created lists of recommendations (items)
2) Evaluation of recommendations order
3) Evaluation of predictions for items bought (if rated)
4) Definition of users with close tastes

1) Evaluate a list of recommendations
2) Evaluate recommendations order
3) Definition of users with close tastes
A system generates predicted ratings $\hat{r}_{ui}$ for a test set $\mathcal{T}$ of user-item pairs $(u, i)$ for which the true ratings $r_{ui}$ are known.

$$\text{RMSE} = \sqrt{\frac{1}{|\mathcal{T}|} \sum_{(u,i) \in \mathcal{T}} (\hat{r}_{ui} - r_{ui})^2}$$

$$\text{MAE} = \sqrt{\frac{1}{|\mathcal{T}|} \sum_{(u,i) \in \mathcal{T}} |\hat{r}_{ui} - r_{ui}|}$$

RMSE disproportionally penalizes large errors: prefers several small errors to one big error.

(Guy Shani and Asela Gunawardana, Evaluating Recommendation Systems)
APPLICABILITY OF THE METRIC

Use case

1) Comparison: real score to the predicted one
2) Evaluation of recommendations order
3) Evaluation of created lists of recommendations (films)
4) Definition of users with close tastes

1) Evaluation on created lists of recommendations (items)
2) Evaluation of recommendations order
3) Comparison: real score to the predicted one (if rated)
4) Definition of users with close tastes

1) Evaluate a list of recommendations
2) Evaluate recommendations order
3) Definition of users with close tastes
4) Comparison: real score to the predicted one
METRICS - PRECISION/RECALL

\[ P = \frac{|\{\text{relevant} \} \cap \{\text{retrieved} \}|}{|\{\text{retrieved} \}|} \]

\[ R = \frac{|\{\text{relevant} \} \cap \{\text{retrieved} \}|}{|\{\text{relevant} \}|} \]

\[
\text{Precision} = \frac{\#tp}{\#tp + \#fp}
\]

\[
\text{Recall (True Positive Rate)} = \frac{\#tp}{\#tp + \#fn}
\]

Interpretation:

Precision= what % of recommended items occurred to be relevant for the user

Recall= what % of the used items is recommended
APPLICABILITY OF THE METRIC

Scenarios

1) Comparison: real score to the predicted one
2) Evaluation of created lists of recommendations (films) precision+/recall ~
   (not sure which recommended items shown to user have been „registered” by the user)
3) Recommendations order
4) Definition of users with close tastes

Platforms

1) Evaluation on created lists of recommendations (items)
2) Recommendations order
3) Evaluation of predictions for items bought (if rated)
4) Definition of users with close tastes

1) Evaluate a list of recommendations
2) Evaluate recommendations order
3) Definition of users with close tastes
**DCG/NDCG (Discounted Cumulative Gain) – Utility of Recommendation**

Assumption: Highly relevant documents are more useful than marginally relevant documents, which are in turn more useful than irrelevant documents.

\[
DCG = \frac{1}{N} \sum_{u=1}^{N} \sum_{j=1}^{J} \frac{g_{uij}}{\max(1, \log_b j)}
\]

\[
NDCG = \frac{DCG}{DCG^*}
\]

where:
- \(N\) = number of users
- \(J\) = item’s count
- \(g_{uij}\) = gain of user \(u\) from using item \(i\)
- \(DCG^*\) = Ideal DCG
**EXAMPLE FOR NDCG**

<table>
<thead>
<tr>
<th>item</th>
<th>score</th>
<th>Max(1;log₂J)</th>
<th>DCG</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>log₂3</td>
<td>3/log₂3</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Sum</td>
<td></td>
<td></td>
<td>5,875</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>item</th>
<th>Score*</th>
<th>Max(1;log₂J)</th>
<th>DCG*</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>log₂3</td>
<td>3/log₂3</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Sum</td>
<td></td>
<td></td>
<td>6,875</td>
</tr>
</tbody>
</table>

\[
\text{NDCG} = \frac{DCG}{DCG^*} = \frac{5,875}{6,875} = 0.85(45)
\]
APPLICABILITY OF THE METRIC

Platforms

Scenarios

1) Comparison: real score to the predicted one
2) Evaluation of created lists of recommendations (films)
3) Evaluation of recommendations order
4) Definition of users with close tastes

1) Evaluation on created lists of recommendations (items)
2) Evaluation of recommendations order
3) Evaluation of predictions for items bought (if rated)
4) Definition of users with close tastes

1) Evaluate a list of recommendations
2) Evaluate recommendations order
3) Definition of users with close tastes
AP/MAP@X (AVERAGE PRECISION/MÉAN AVERAGE PRECISION)

MAP is just an average of APs, or average precision, for all users.
recommend at most x items for each user
order matters

Formula: \( AP(x) = \sum_{i=1}^{x} (\text{precision at } i \times \text{change in recall at } i) \)

Precision at \( i \) is a percentage of correct items among first \( i \) recommendations.
Change in recall is \( 1/x \) if item at \( i \) is correct (for every correct item), otherwise zero.
Example: \([1,2,3,4,5]\) is the correct order, \([6,4,7,1,2]\) is recommended
\( AP@2 = 0 \times 0 + 0.5 \times 0.5 = 0.25 \)
APPLICABILITY OF THE METRIC

Plattforms

- moviepilot
- Amazon
- YouTube

Scenarios

1) Comparison: real score to the predicted one
2) Evaluation of created lists of recommendations (films)
3) Evaluation of recommendations order
4) Definition of users with close tastes

1) Evaluation on created lists of recommendations (items)
2) Evaluation of recommendations order
3) Evaluation of predictions for items bought (if rated)
4) Definition of users with close tastes

1) Evaluate a list of recommendations
2) Evaluate recommendations order
3) Definition of users with close tastes
EXPECTED RECIPROCAL RANK

\[
ERR := \sum_{r=1}^{n} \frac{1}{r} \prod_{i=1}^{r-1} (1 - R_i) R_r.
\]

Number of documents in the ranking

Probability of user to be satisfied with document found on position before \( r \)

Probability of user to be satisfied with document found on position \( r \)

Probability of user to stop searching at position \( r \)

Gain

Mapping function from Relevance of grades to Probability of relevance

\[
R_i := \mathcal{R}(g_i),
\]

\[
\mathcal{R}(g) := \frac{2^g - 1}{2^g_{\text{max}}}, \quad g \in \{0, \ldots, g_{\text{max}}\}.
\]

with grades
- Perfect
- Excellent
- Good
- Fair
- Bad

Are respectively \( \approx 0.96; 0.46; 0.22; 0.09; 0. \)

ERR implicitly discounts documents based on the relevance of previously seen documents.
APPLICABILITY OF THE METRIC

Plattforms

Scenarios

1) Comparison: real score to the predicted one
2) Evaluation of created lists of recommendations (films)
3) Evaluation of recommendations order
4) Definition of users with close tastes

1) Evaluation on created lists of recommendations (items)
2) Evaluation of recommendations order
3) Evaluation of predictions for items bought (if rated)
4) Definition of users with close tastes

1) Evaluate a list of recommendations
2) Evaluate recommendations order
3) Definition of users with close tastes
PEARSON CORRELATION

\[ \rho_{X,Y} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y} \]

\[ \mathcal{P}(u, v) = \frac{\sum_{i \in I_{uv}} (r_{ui} - \bar{r}_u)(r_{vi} - \bar{r}_v)}{\sqrt{\sum_{i \in I_{uv}} (r_{ui} - \bar{r}_u)^2} \sqrt{\sum_{i \in I_{uv}} (r_{vi} - \bar{r}_v)^2}} \]

Used for: search of KNN in memory-based collaborative filtering models

intersection of user u's and v's rated items

the rating score of user u for item i

average score of user u
APPLICABILITY OF THE METRIC

Plattforms

Scenarios

1) Comparison: real score to the predicted one
2) Evaluation of created lists of recommendations (films)
3) Evaluation of recommendations order
4) Definition of users with close tastes

1) Evaluation on created lists of recommendations (items)
2) Evaluation of recommendations order
3) Evaluation of predictions for items bought (if rated)
4) Definition of users with close tastes

1) Evaluate a list of recommendations
2) Evaluate recommendations order
3) Definition of users with close tastes
AGENDA

Definition
Introduction
Factors defining recommender systems
Properties of recommender systems
Evaluating algorithms
Offline experiment
User study experiment
Online experiment
Metrics to measure prediction accuracy and their applications

Conclusion

Bibliography
Metrics that are defined to handle binary relevance data are not directly suitable for graded relevance data.

- We lose grading information within the rated items
- The choice of the thresholding relevance is arbitrary and will have an impact on the performance of different recommendation approaches


