Web Mining: Recommender Systems: Evaluation

Alípio Jorge, DCC-FC, Universidade do Porto
amjorge@fc.up.pt

What is there to evaluate?

Quality dimensions

- Predictive power
  - Guess what users will like next
  - Tendency to inertia
- Novelty
  - Things I didn't know about
- Serendipity
  - Things I would very hardly find

Quality dimensions

- Diversity
  - Not always the same artists!
- Safety / Robustness
  - No other users are tampering with the recommendations
- Privacy preserving
  - Can other users infer my preferences?

Quality dimensions – owner's view

- More sales
  - That's what really matters
- Better sales
  - Sell what you want to sell
- Loyalty
  - Abandon rate
    - Gone client buys no product
- Reputation
  - Clients value the recommendations and talk about them

Evaluating recommender models/systems

- How can we measure the success of a recommender?

- Offline evaluation
  - Cheap, repeatable
  - Not the real thing, no user feedback

Evaluating recommender models/systems

- Online evaluation
  - User interacts
  - More expensive, interferes with business, not repeatable

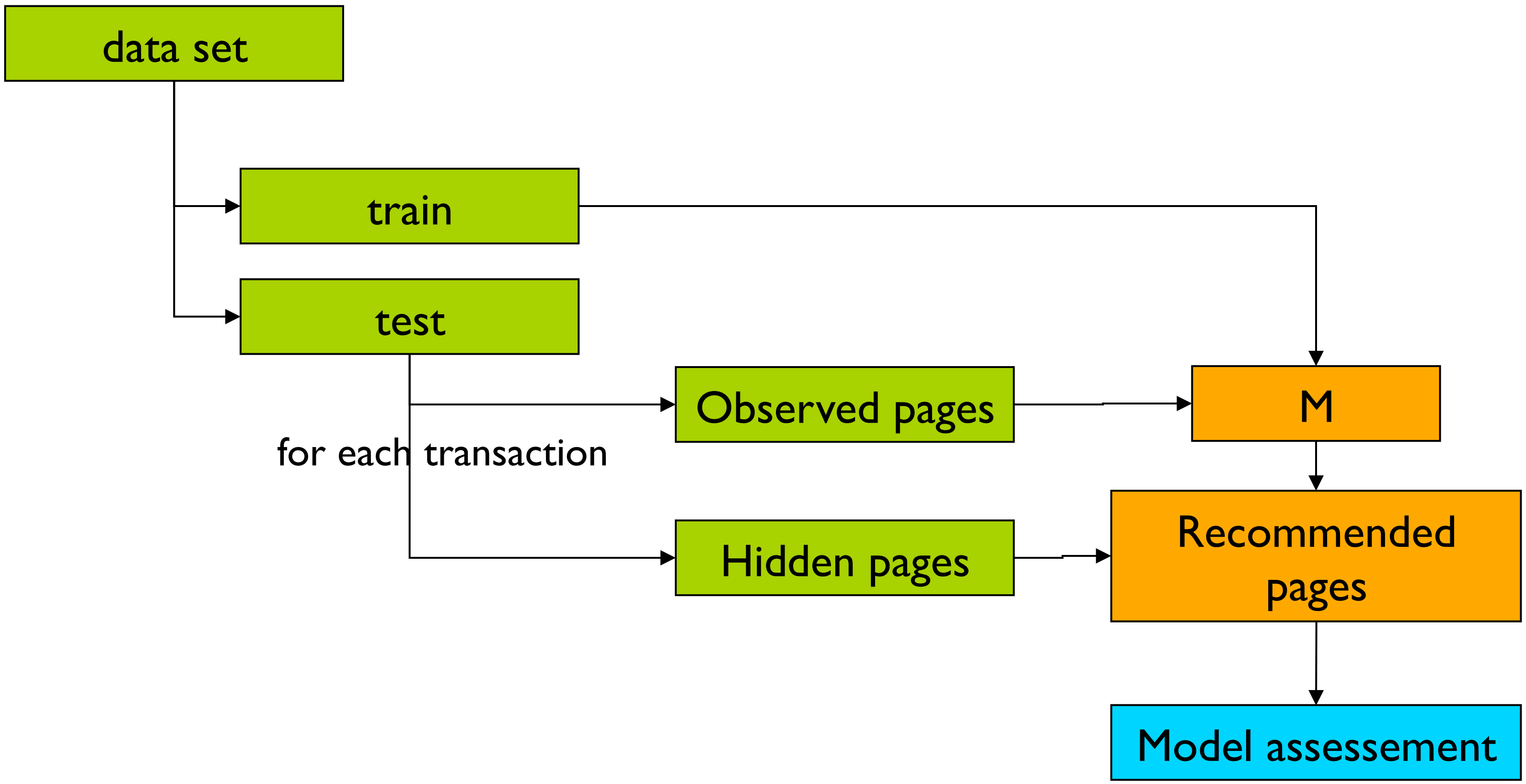
Evaluating recommender models/systems

- User studies
  - User behavior, qualitative feedback
  - Expensive, limited samples

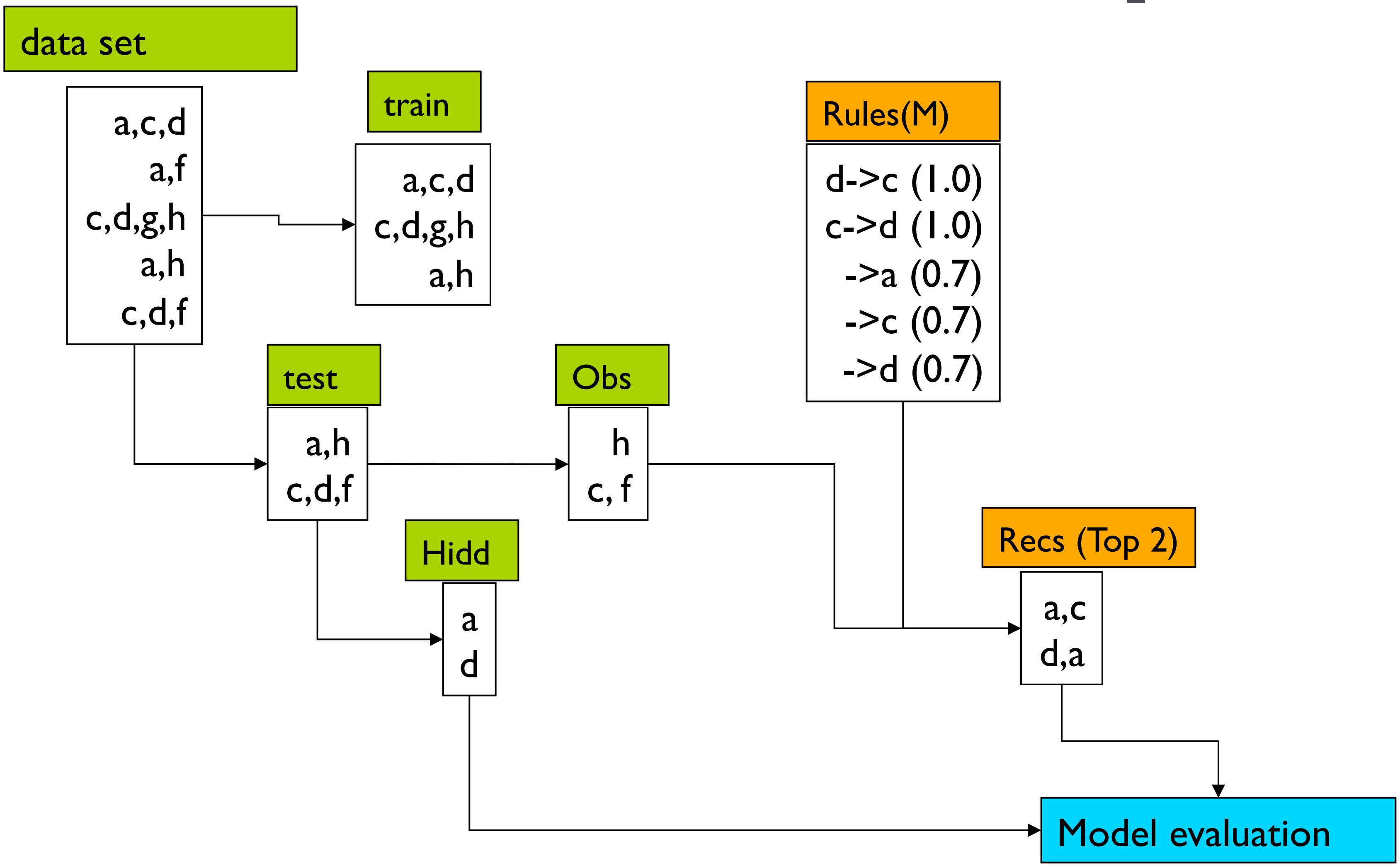
Offline Evaluation

Offline evaluation: All-but-one proc.

- Treino/teste e o protocolo “all-but-one”



Offline evaluation: All-but-one proc.



Measures for item recommendation (top-K)

Success measures

Recall: percentage of relevant items guessed

$$\text{Recall} = \frac{\#\left(\text{Hidden} \cap \text{Recommended}\right)}{\#\,\text{Hidden}}$$

Precision: average quality of each recommendation

$$\text{Precision} = \frac{\#\left(\text{Hidden} \cap \text{Recommended}\right)}{\#\,\text{Recommended}}$$

F1: combines recall and precision (harmonic mean)

$$\text{F1} = \frac{2 \times recall \times precision}{recall + precision}$$

amjorge@fc.up.pt

Other success measures – top K

- Other measures
  - Recall@K
  - Precision@K
  - MAP – mean average precision
    - AP@K - average of P@1, P@2 .. P@K
    - MAP – average AP over N users
  - NDCG – Normalized Discounted Cumulative Gain
    - consider K recommendations
    - fading sum of relevance of recommendations – DCG
    - divide by ideal DCG
  - Any ranking measure
  - …

Measures for rating predictions

Success measures (ratings)

RMSE: root mean squared error

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

MAE: mean average error

$$\mathrm{MAE} = \frac{1}{|\mathrm{T}|}\sum_{(u,i)\in\mathrm{T}}\left|\hat{r}_{ui} - r_{ui}\right|$$

amjorge@fc.up.pt

More info

Micro and macro averaging

- Example
  - Hidden: ab, c, ac
  - Rec: ac, ac, a
- Calculate Recall, Precision, F1
  - **micro** and **macro** averaging
    - Micro: average of small parts
    - Macro: average of large parts (each user has same weight)

Hid	Recs	#hits	#hid	#recs	Recall	Prec	F1
a,b	a,c	1	2	2	?	?	?
c	a,c	1	1	2	?	?	?
a,c	a	1	2	1	?	?	?

Micro and macro averaging

- Example
  - Hidden: ab, c, ac
  - Rec: ac, ac, a

- Calculate Recall, Precision, F1
  - **micro** and **macro** averaging
    - Micro: average of small parts
    - Macro: average of large parts

Hid	Recs	#hits	#hid	#recs	Recall	Prec	F1
a,b	a,c	1	2	2	?	?	?
c	a,c	1	1	2	?	?	?
a,c	a	1	2	1	?	?	?

amjorge@fc.up.pt

Other evaluation procedures

- What to guess
  - Try to guess last item of each test session
  - Try to guess each item in the session from previous ones
- Train / test
  - older sessions to train, newer to test
  - sliding window
  - growing window

Resources

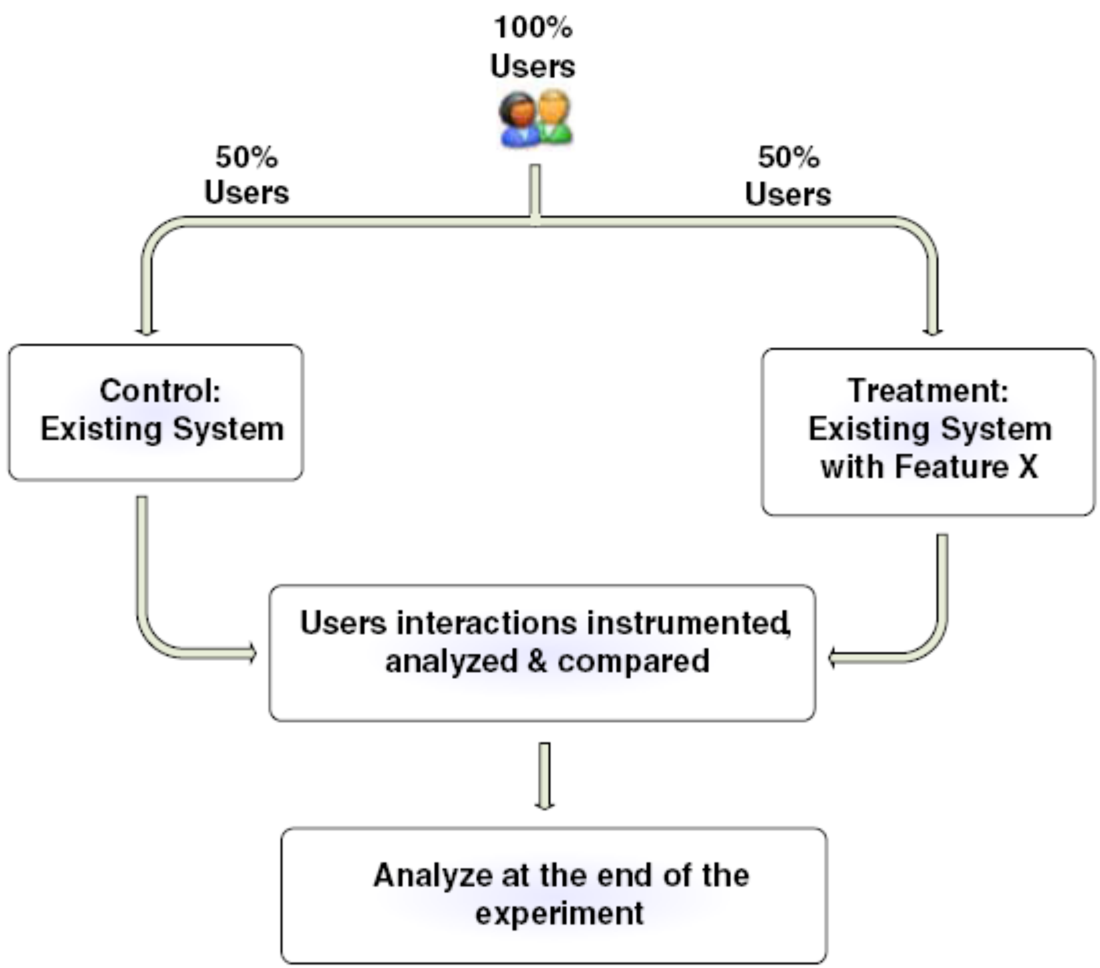
- Articles
  - J. Breese, D. Heckerman, C. Kadie, and others, "Empirical analysis of predictive algorithms for collaborative filtering," Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, vol. 461, 1998.

Online Evaluation

Why?

- w.r.t. offline experiments
  - the real thing, the real time
  - automation interferes with the site
  - need to continuously monitor
- w.r.t. to the HIPPO (highest paid person's opinion)
  - listen to the users!
  - great ideas may not be so great

Controlled experiments (A/B tests)



(Kohavi 2009)

Controlled experiments - verbum

- Success metrics
  - overall evaluation criterion is a good idea
- Factors / variables
  - what we want to try
    - with / without "treatment"
    - number of items on the screen, etc.
- Experimental Unit
  - subjects of test: typically users
  - what goes to control? What goes to treatment?

Controlled experiments - verbum

- Null hypothesis H0
  - with treatment or without is the same
- Confidence level
  - Maximum admissiible probability of observing extreme values of a given statistics when H0 is true - P(acc H0 | H0)
- Power
  - probability of correctly rejecting H0 – P(rej H0 | ~H0)
- A/A test
  - placebo
  - H0 should be rejected 5% of the times if confidence level is 95%

Issues

- Sample size
- Proportion
  - typically 50-50
- Robots
- Treatment ramp-up
  - monotonic ramp-up
  - automatic
  - abort if bug
- Automation

Implementation

- randomization
  - each user is randomly assigned to one group
  - once assigned should stick to that group (consistency)
- time
  - a new feature may fail because it is too slow
- where/how to split
  - proxy / server

Resources

- Book
  - Web Data Mining, Bing Liu
- Articles
  - Ron Kohavi, Roger Longbotham, Dan Sommerfield, Randal M. Henne: Controlled experiments on the web: survey and practical guide. Data Min. Knowl. Discov. 18(1): 140-181 (2009)
  - Ron Kohavi, Randal M. Henne, Dan Sommerfield: Practical guide to controlled experiments on the web: listen to your customers not to the hippo. KDD 2007: 959-967
  - google "DBLP Ron Kohavi"