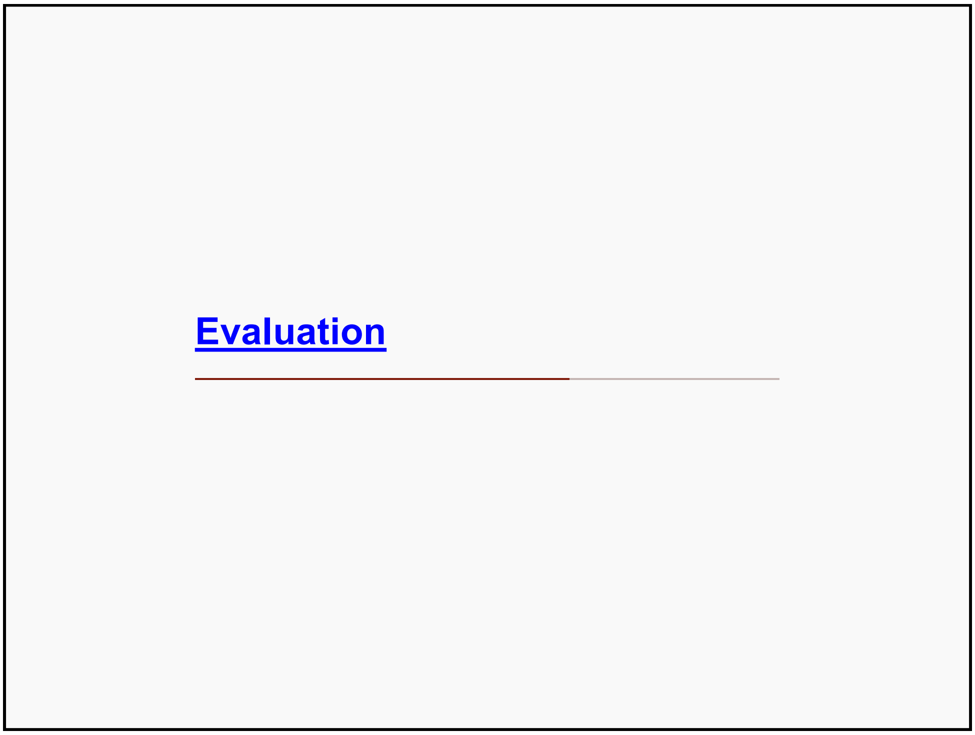
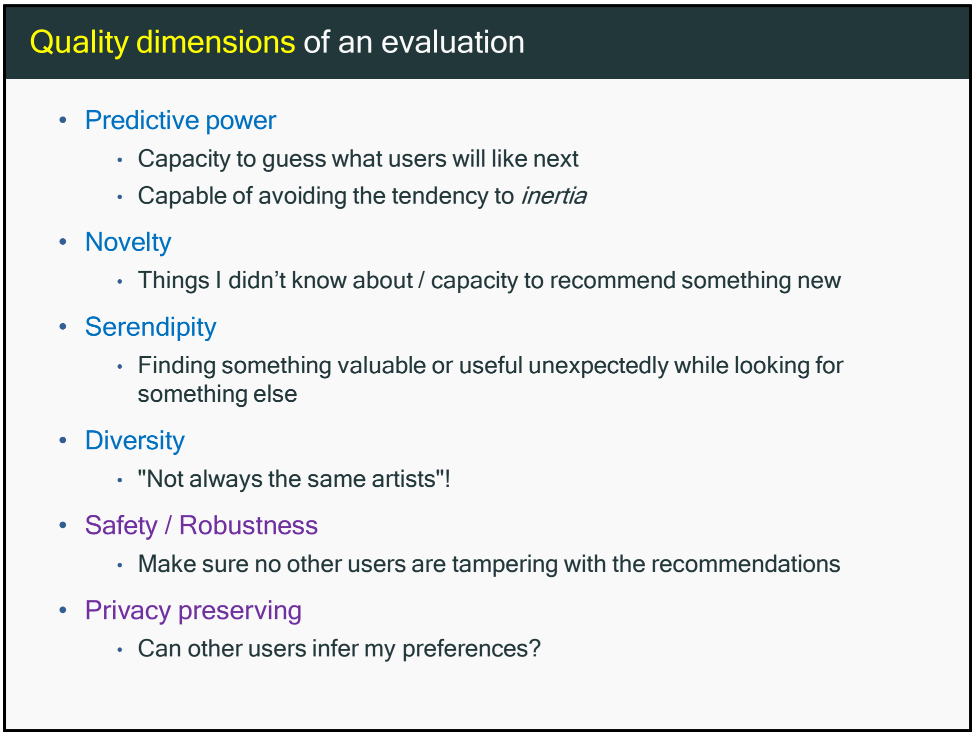


Evaluation

58



Quality dimensions of an evaluation

- Predictive power
  - Capacity to guess what users will like next
  - Capable of avoiding the tendency to *inertia*
- Novelty
  - Things I didn’t know about / capacity to recommend something new
- Serendipity
  - Finding something valuable or useful unexpectedly while looking for something else
- Diversity
  - "Not always the same artists"!
- Safety / Robustness
  - Make sure no other users are tampering with the recommendations
- Privacy preserving
  - Can other users infer my preferences?

59

Quality evaluation - the Owner’s view

- More sales
  - That’s what really matters!
- Better sales
  - Sell what you want to sell
- Improve loyalty
  - Reduce abandon rate
    - because, gone clients buy no products
- Reputation
  - Clients value the recommendations and talk about them

60

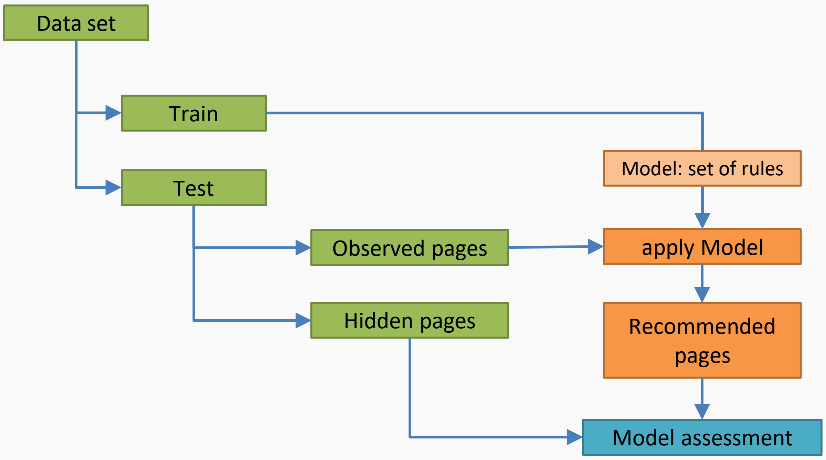
Recommender Systems' evaluation

- How can we measure the success of a recommender model?
- Offline Evaluation
  - Cheap, repeatable
  - Not the real thing, no user feedback
- Online Evaluation
  - User interacts (leading to "active learning")
  - More expensive, interferes with business, not repeatable
- User Studies
  - User behavior, qualitative feedback
  - Expensive, limited samples

61

Offline evaluation

Standard protocol:



Data set

Train

Test

Model: set of rules

Observed pages

apply Model

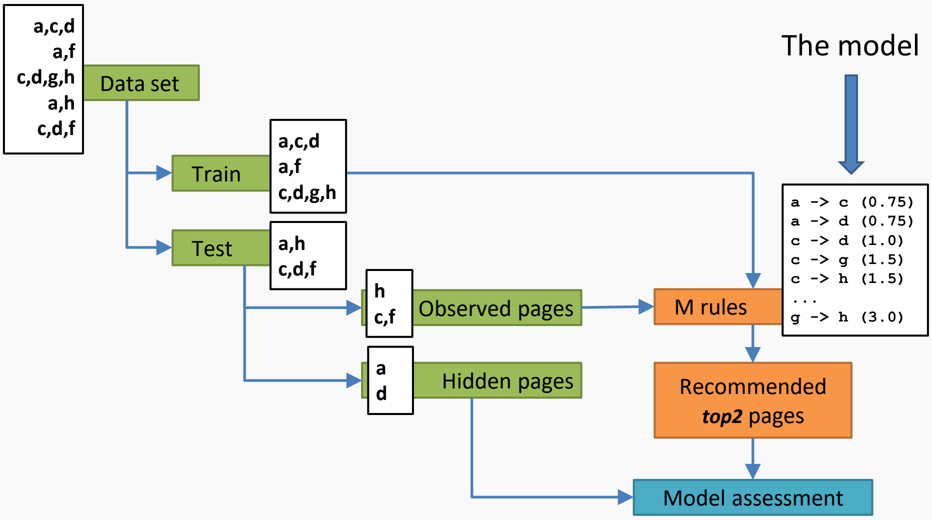
Hidden pages

Recommended pages

Model assessment

62

Offline evaluation (cont.)



a,c,d
a,f
c,d,g,h
a,h
c,d,f

Data set

Train

a,c,d
a,f
c,d,g,h

Test

a,h
c,d,f

h
c,f

Observed pages

a
d

Hidden pages

The model

M rules

```
a -> c (0.75)
a -> d (0.75)
c -> d (1.0)
c -> g (1.5)
c -> h (1.5)
...
g -> h (3.0)
```

Recommended *top2* pages

Model assessment

63

Success metrics for top-k item recommendation

- Recall: proportion of relevant items that were recommended

$$\textbf{\textit{recall}} = \frac{\#(Hidden \cap Recommended)}{\#Hidden} = \frac{\#Hits}{\#Hidden}$$

- Precision: proportion of recommended items that are relevant

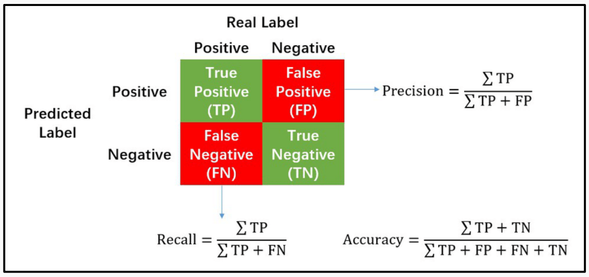
$$\textbf{\textit{precision}} = \frac{\#(Hidden \cap Recommended)}{\#Recommended} = \frac{\#Hits}{\#Recommended}$$

- F1: the trade-off between relevance and completeness of recommend items. It is the harmonic mean of precision and recall.

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

64

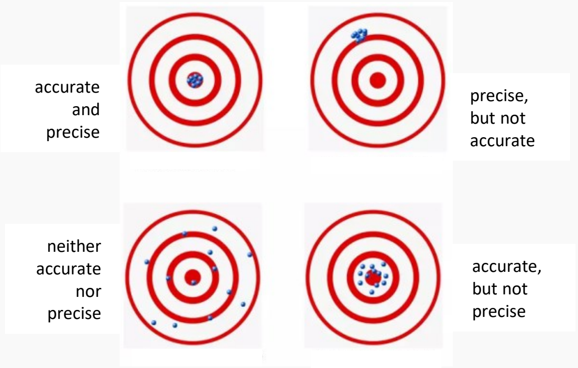
Visually understanding evaluation metrics

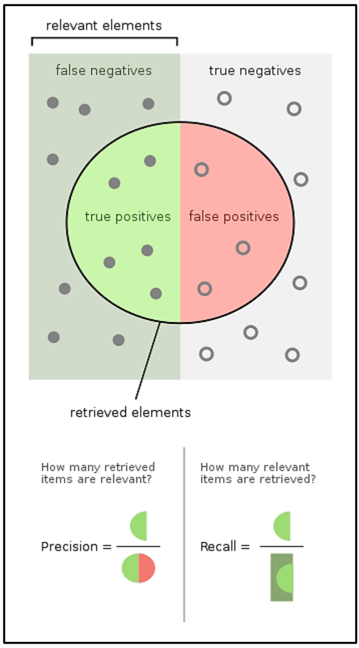


		Real Label	
		Positive	Negative
Predicted Label	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

$$\text{Precision} = \frac{\sum \text{TP}}{\sum \text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\sum \text{TP}}{\sum \text{TP} + \text{FN}} \qquad \text{Accuracy} = \frac{\sum \text{TP} + \text{TN}}{\sum \text{TP} + \text{FP} + \text{FN} + \text{TN}}$$


accurate and precise
precise, but not accurate
neither accurate nor precise
accurate, but not precise



65

Success metrics for recommendation with ratings

We are assessing the rating error:

- RMSE (Root Mean Squared Error)

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

- MAE (Mean Absolute Error)

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

Obs: RMSE is more sensitive to large errors compared to other metrics like MAE

Note: *T* is the set of transactions

66

Other evaluation procedures for session-based recommenders

- What to guess?

  How the RS should make predictions in sessions

  - try to guess the last item in each test session
  - try to guess each item in the same session from previous interacted items

- Train / test:

  How to split the dataset into train and test subsets

  - older sessions to train, newer to test
  - fixed-size sliding window for training and following session for testing
  - growing training window over time

67

Online evaluation

Online evaluation, also known as A/B testing or live testing, refers to testing the **effectiveness of a system in a real-world setting**, with actual users and real-time interactions.

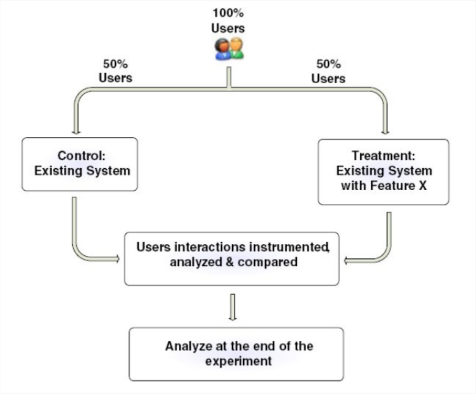
Why?

1. **Influence user behavior**: Online evaluation helps measure the impact of changes on user engagement and other KPIs
2. **Real-world testing**: Provides insights into system performance under real-world conditions. The evolving user preferences and interactions could not be captured in offline evaluation
3. **Automation interference**: Continuously monitoring during online evaluation helps address issues arising from offline (automatic) recommendations
4. **Challenging HIPPO**: Online evaluation promotes data-driven decisions instead of relying solely on the highest-paid person's opinion
5. **Listening to users**: Online evaluation gathers direct user feedback for improvements in the system or user interface
6. **Validating ideas**: Online evaluation tests the true effectiveness of ideas with actual users, ensuring data-driven implementation decisions

68

Online evaluation (cont.)

- **Goal:** measure the change in user behavior when interacting with different recommendation systems.
- **Strategy:** controlled experiments (A/B tests)



100% Users

50% Users

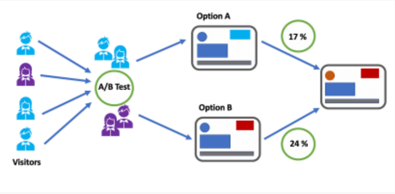
50% Users

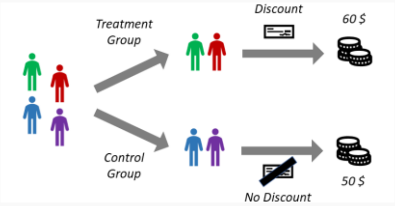
Control: Existing System

Treatment: Existing System with Feature X

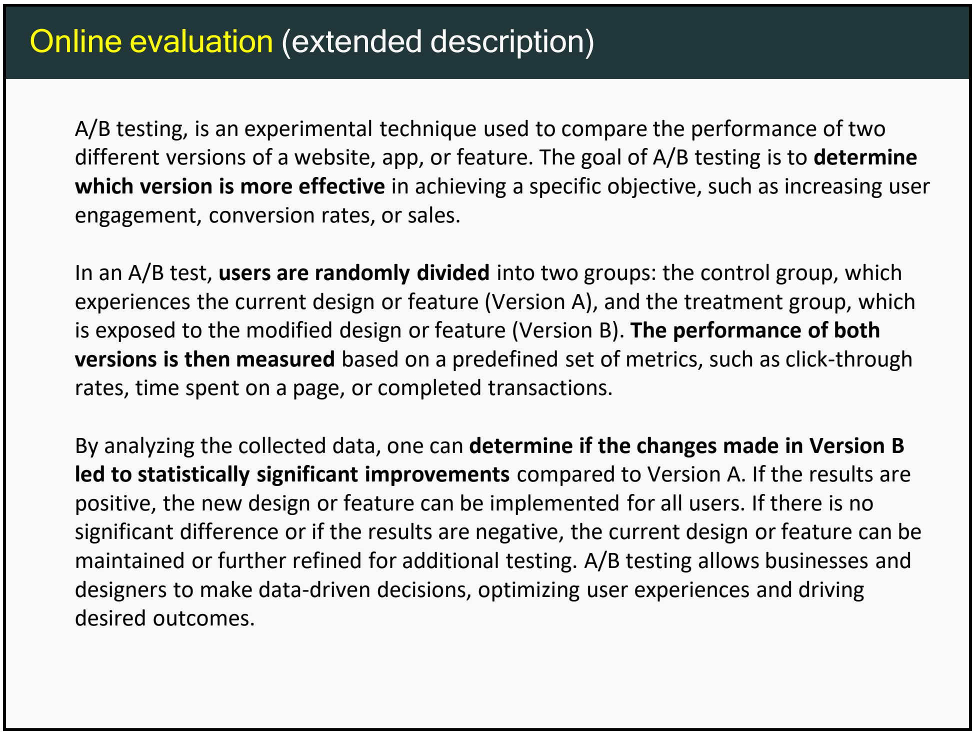
Users interactions instrumented, analyzed & compared

Analyze at the end of the experiment





69



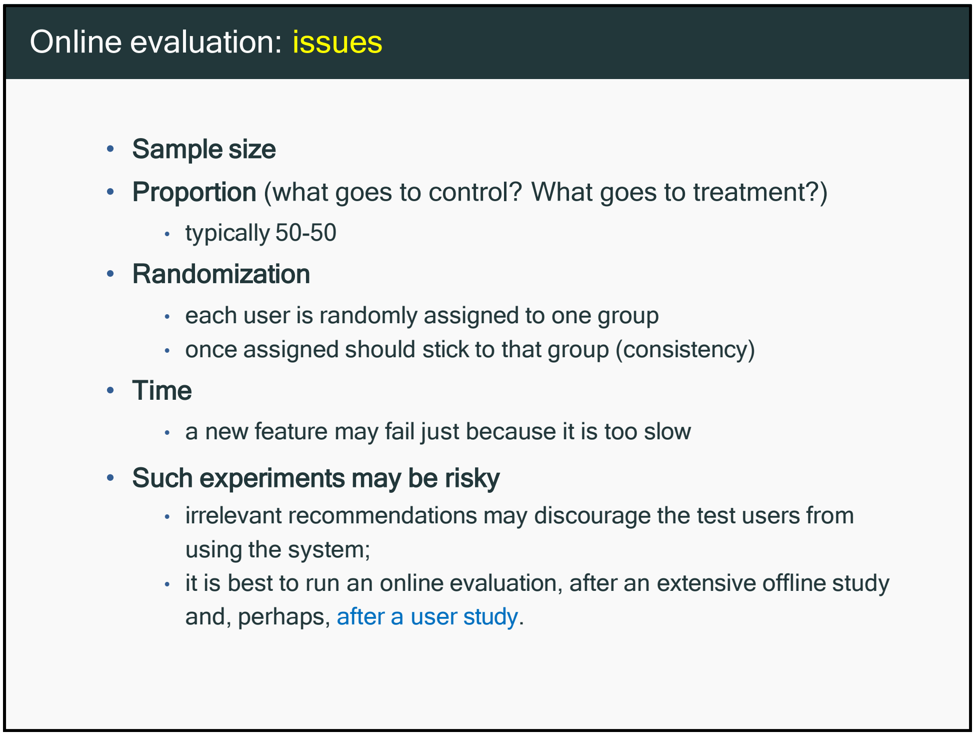
Online evaluation (extended description)

A/B testing, is an experimental technique used to compare the performance of two different versions of a website, app, or feature. The goal of A/B testing is to **determine which version is more effective** in achieving a specific objective, such as increasing user engagement, conversion rates, or sales.

In an A/B test, **users are randomly divided** into two groups: the control group, which experiences the current design or feature (Version A), and the treatment group, which is exposed to the modified design or feature (Version B). **The performance of both versions is then measured** based on a predefined set of metrics, such as click-through rates, time spent on a page, or completed transactions.

By analyzing the collected data, one can **determine if the changes made in Version B led to statistically significant improvements** compared to Version A. If the results are positive, the new design or feature can be implemented for all users. If there is no significant difference or if the results are negative, the current design or feature can be maintained or further refined for additional testing. A/B testing allows businesses and designers to make data-driven decisions, optimizing user experiences and driving desired outcomes.

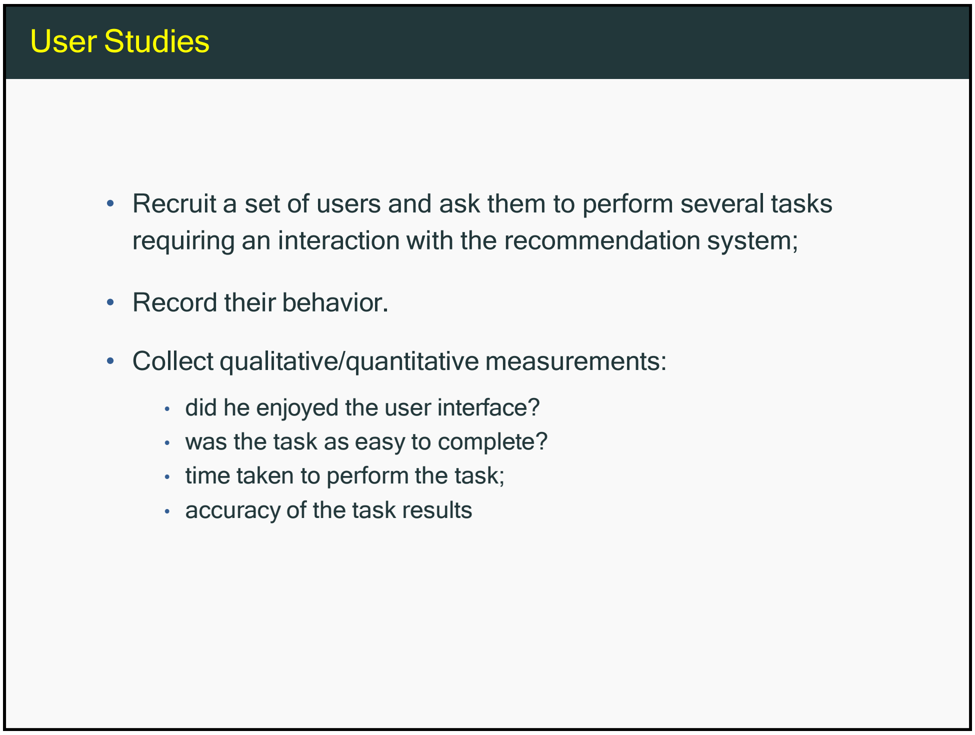
70



Online evaluation: issues

- Sample size
- Proportion (what goes to control? What goes to treatment?)
  - typically 50-50
- Randomization
  - each user is randomly assigned to one group
  - once assigned should stick to that group (consistency)
- Time
  - a new feature may fail just because it is too slow
- Such experiments may be risky
  - irrelevant recommendations may discourage the test users from using the system;
  - it is best to run an online evaluation, after an extensive offline study and, perhaps, after a user study.

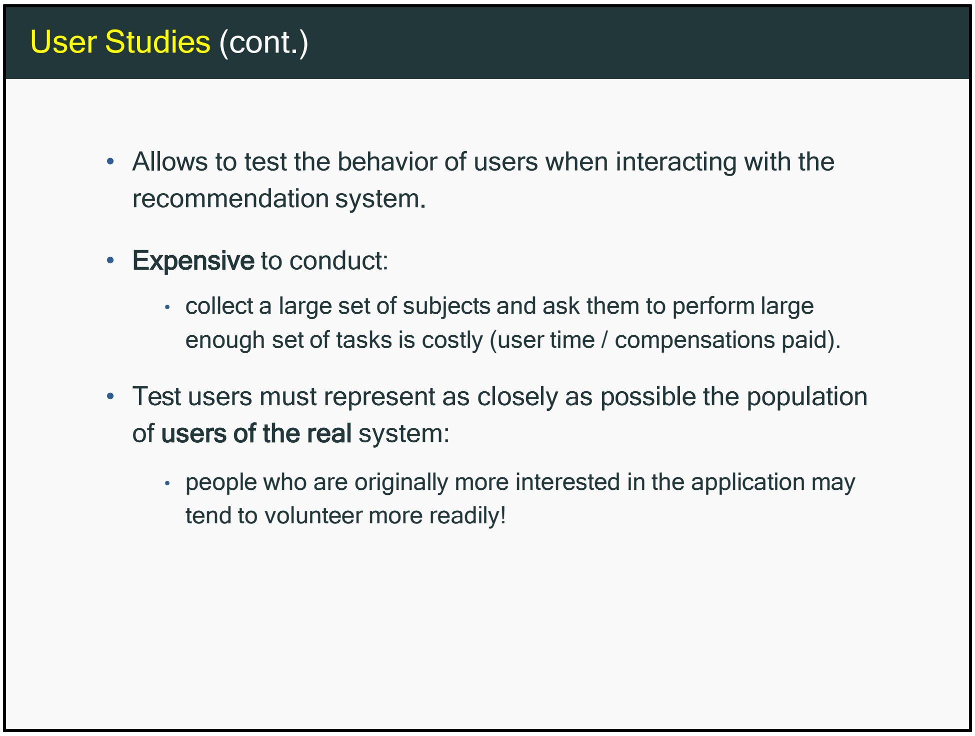
71



User Studies

- Recruit a set of users and ask them to perform several tasks requiring an interaction with the recommendation system;
- Record their behavior.
- Collect qualitative/quantitative measurements:
  - did he enjoyed the user interface?
  - was the task as easy to complete?
  - time taken to perform the task;
  - accuracy of the task results

72



User Studies (cont.)

- Allows to test the behavior of users when interacting with the recommendation system.
- **Expensive** to conduct:
  - collect a large set of subjects and ask them to perform large enough set of tasks is costly (user time / compensations paid).
- Test users must represent as closely as possible the population of **users of the real** system:
  - people who are originally more interested in the application may tend to volunteer more readily!

73

Other success metrics for top-k items recommendations

- Recall@k = (number of relevant items retrieved **in top k**) / (total number of relevant items)
- Precision@k =(number of relevant items retrieved **in top k**) / **k**
- MAP = Mean Average Precision (for ranking of results)
- AP_u@k = *Sum*(Precision@i × relev_i) / *min*(k, R_u)
  - Precision@i is the precision at the i-th recommendation
  - relev_i is a boolean function to whether the item at i is relevant to the user
  - R_u is the total number of relevant items for user u
- MAP@k = *Sum*(AP_u@k) / |U| ; |U| the number of users
- NDCG = Normalized Discounted Cumulative Gain
  - consider K recommendations
  - fading sum of relevance of recommendations - DCG
  - divide by ideal DCG
- ...

74

Micro and Macro Averaging of measures

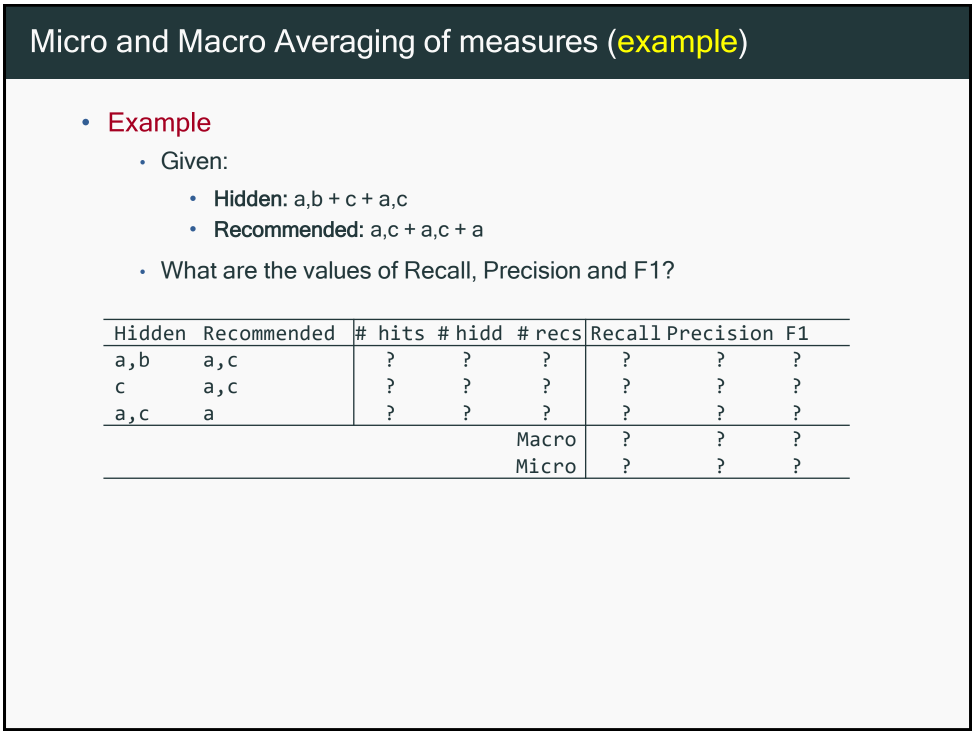
Micro Averaging: **aggregate performance of different samples** treating them independently of the size

- is calculated by summing the results for all samples, rather than calculating them for each sample separately
- Should be taken care when there is a class imbalance

Macro Averaging: aggregates performance across different samples, **calculating performance for each sample** separately and then the average across all samples.

- each sample has the same weight regardless of the size of the sample.
- this can be useful when focusing on performance of each sample/class
- Possible to calculate precision, recall, and F1 for each class/sample

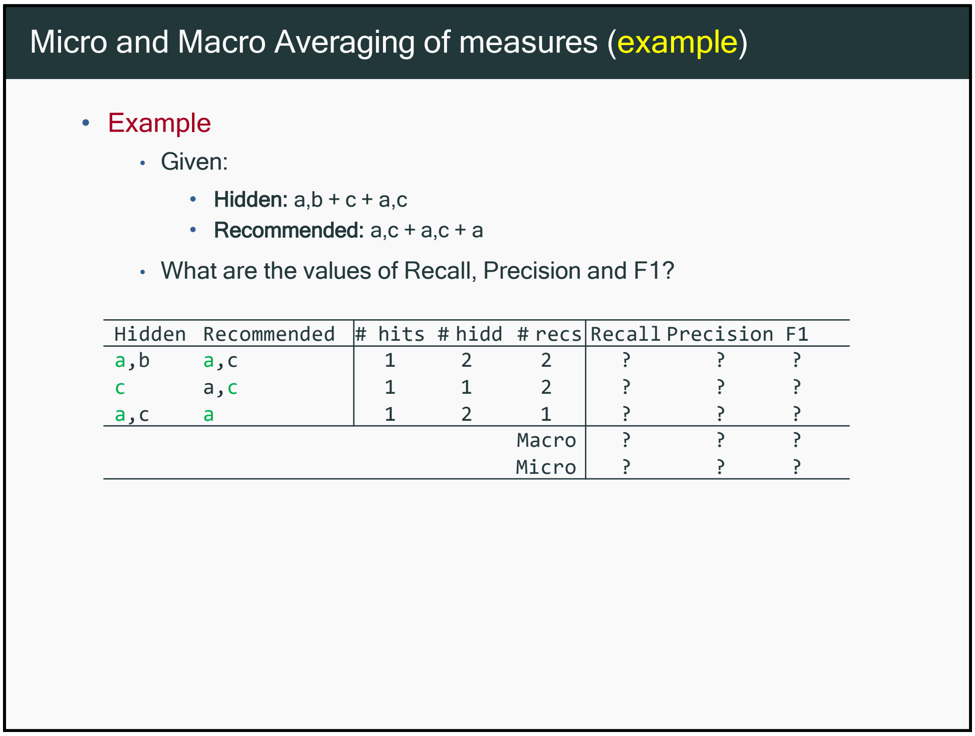
75



Micro and Macro Averaging of measures (example)

- Example
  - Given:
    - **Hidden:** a,b + c + a,c
    - **Recommended:** a,c + a,c + a
  - What are the values of Recall, Precision and F1?

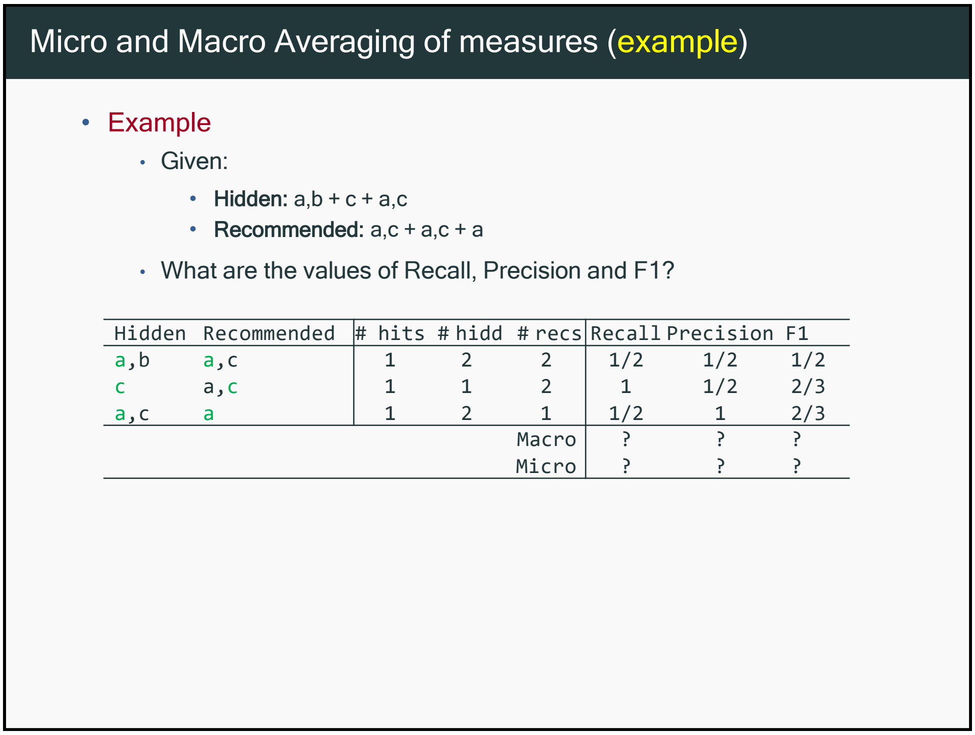
Hidden	Recommended	# hits	# hidd	# recs	Recall	Precision	F1
a,b	a,c	?	?	?	?	?	?
c	a,c	?	?	?	?	?	?
a,c	a	?	?	?	?	?	?
				Macro	?	?	?
				Micro	?	?	?



Micro and Macro Averaging of measures (example)

- Example
  - Given:
    - **Hidden:** a,b + c + a,c
    - **Recommended:** a,c + a,c + a
  - What are the values of Recall, Precision and F1?

Hidden	Recommended	# hits	# hidd	# recs	Recall	Precision	F1
a,b	a,c	1	2	2	?	?	?
c	a,c	1	1	2	?	?	?
a,c	a	1	2	1	?	?	?
				Macro	?	?	?
				Micro	?	?	?

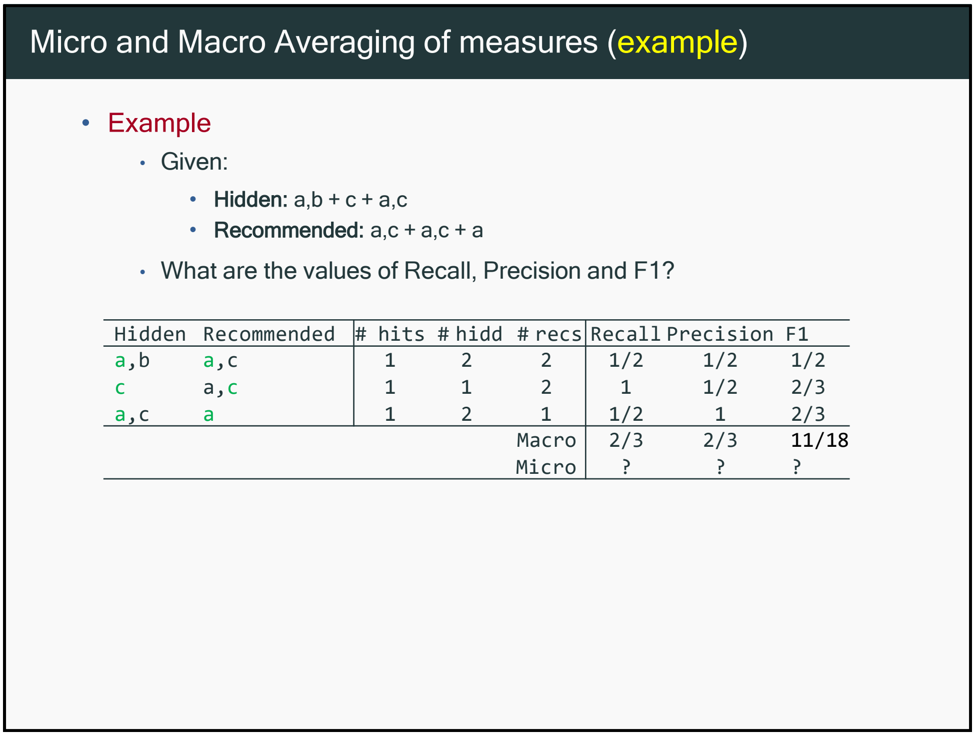


Micro and Macro Averaging of measures (example)

- Example
  - Given:
    - **Hidden:** a,b + c + a,c
    - **Recommended:** a,c + a,c + a
  - What are the values of Recall, Precision and F1?

Hidden	Recommended	# hits	# hidd	# recs	Recall	Precision	F1
a,b	a,c	1	2	2	1/2	1/2	1/2
c	a,c	1	1	2	1	1/2	2/3
a,c	a	1	2	1	1/2	1	2/3
				Macro	?	?	?
				Micro	?	?	?

78

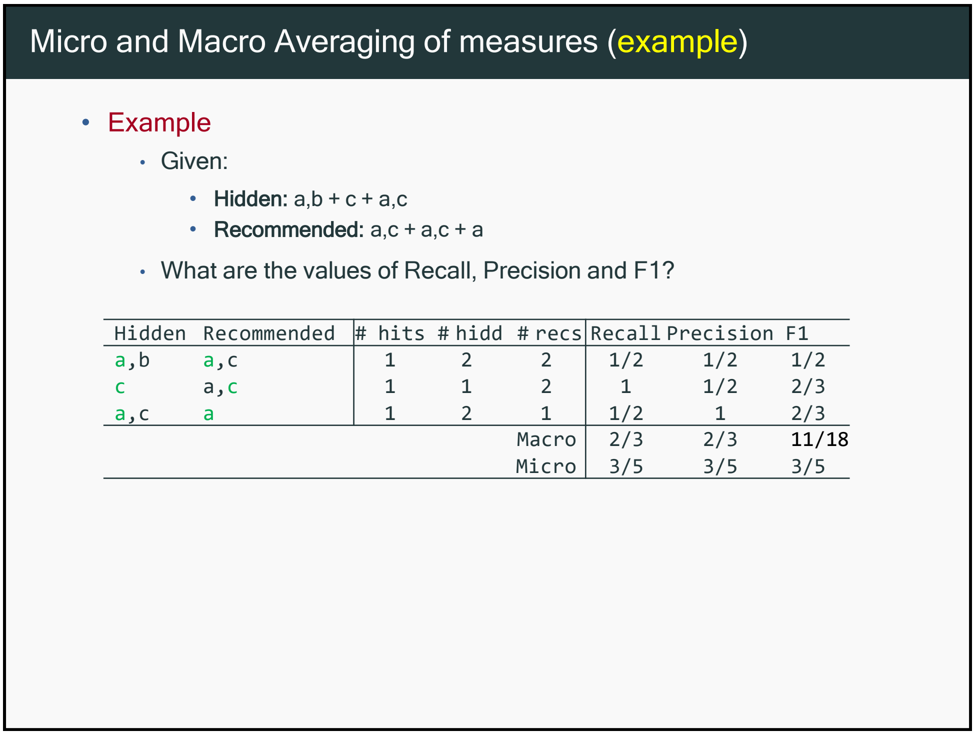


Micro and Macro Averaging of measures (example)

- Example
  - Given:
    - **Hidden:** a,b + c + a,c
    - **Recommended:** a,c + a,c + a
  - What are the values of Recall, Precision and F1?

Hidden	Recommended	# hits	# hidd	# recs	Recall	Precision	F1
a,b	a,c	1	2	2	1/2	1/2	1/2
c	a,c	1	1	2	1	1/2	2/3
a,c	a	1	2	1	1/2	1	2/3
				Macro	2/3	2/3	11/18
				Micro	?	?	?

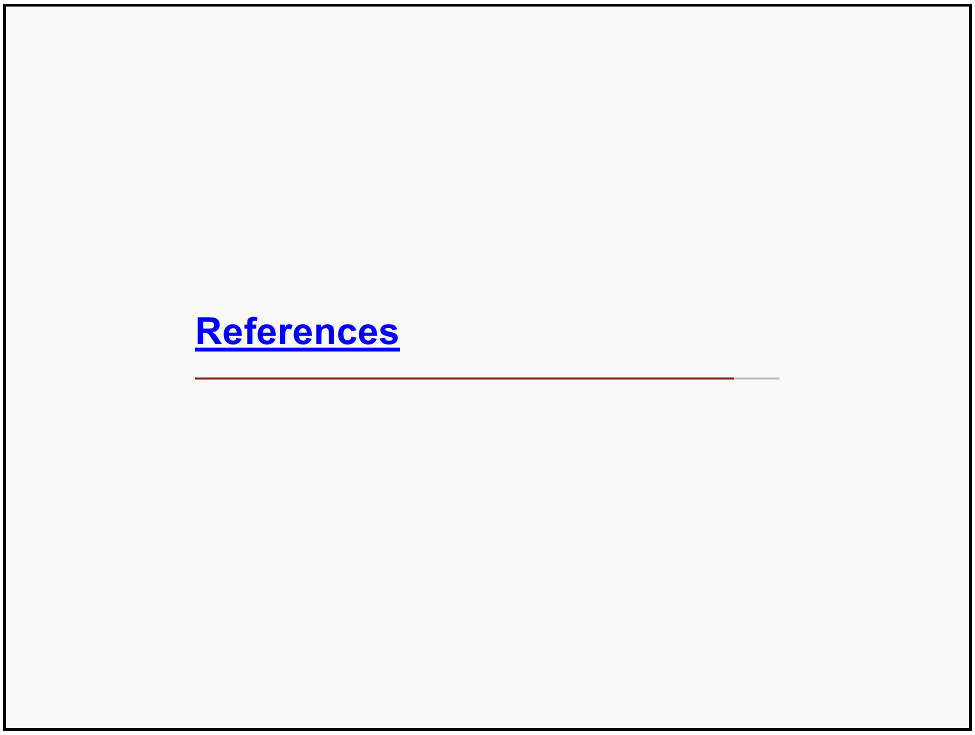
79



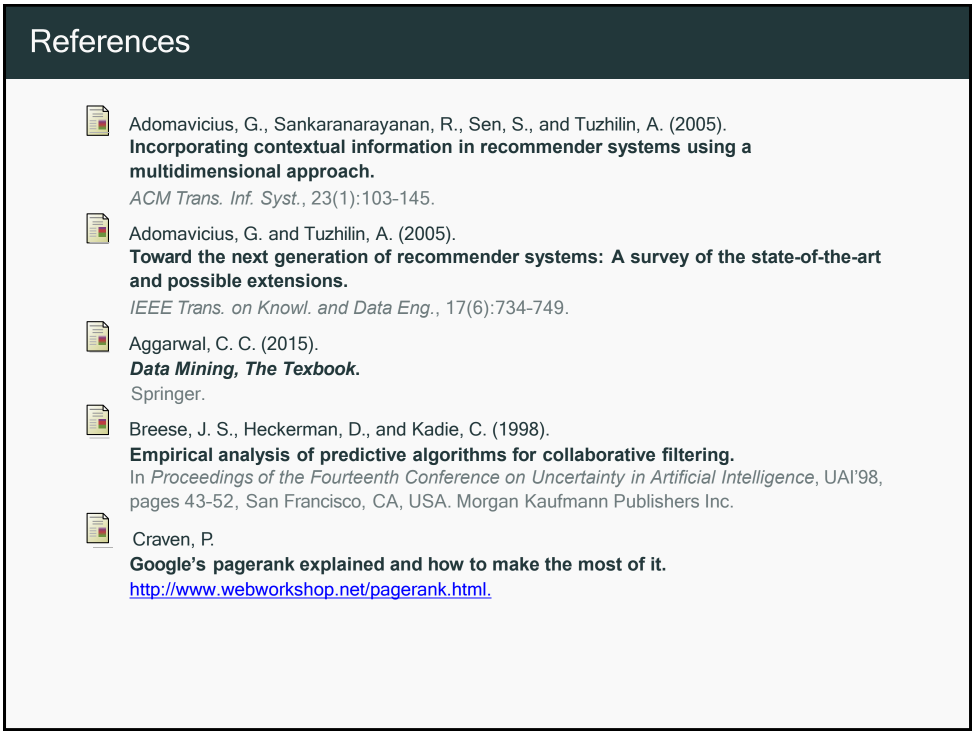
Micro and Macro Averaging of measures (example)

- Example
  - Given:
    - **Hidden:** a,b + c + a,c
    - **Recommended:** a,c + a,c + a
  - What are the values of Recall, Precision and F1?

Hidden	Recommended	# hits	# hidd	# recs	Recall	Precision	F1
a,b	a,c	1	2	2	1/2	1/2	1/2
c	a,c	1	1	2	1	1/2	2/3
a,c	a	1	2	1	1/2	1	2/3
				Macro	2/3	2/3	11/18
				Micro	3/5	3/5	3/5



References



References

Adomavicius, G., Sankaranarayanan, R., Sen, S., and Tuzhilin, A. (2005).
Incorporating contextual information in recommender systems using a multidimensional approach.
ACM Trans. Inf. Syst., 23(1):103-145.

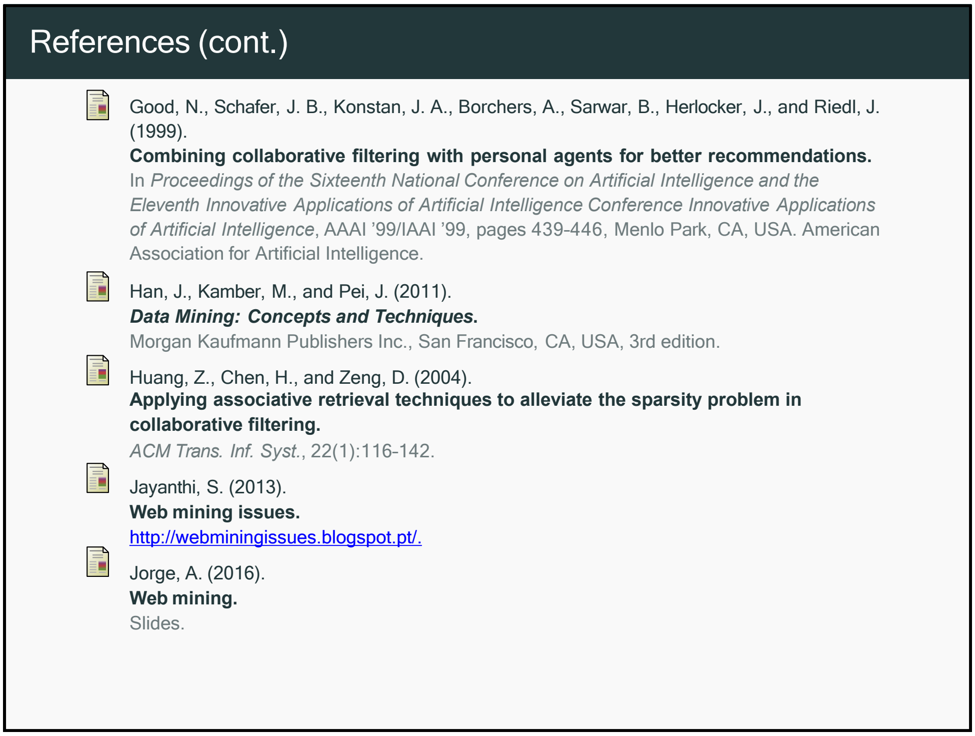
Adomavicius, G. and Tuzhilin, A. (2005).
Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions.
IEEE Trans. on Knowl. and Data Eng., 17(6):734-749.

Aggarwal, C. C. (2015).
Data Mining, The Texbook.
Springer.

Breese, J. S., Heckerman, D., and Kadie, C. (1998).
Empirical analysis of predictive algorithms for collaborative filtering.
In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, UAI'98, pages 43-52, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Craven, P.
Google's pagerank explained and how to make the most of it.
http://www.webworkshop.net/pagerank.html.

82



References (cont.)

Good, N., Schafer, J. B., Konstan, J. A., Borchers, A., Sarwar, B., Herlocker, J., and Riedl, J. (1999).
Combining collaborative filtering with personal agents for better recommendations.
In *Proceedings of the Sixteenth National Conference on Artificial Intelligence and the Eleventh Innovative Applications of Artificial Intelligence Conference Innovative Applications of Artificial Intelligence*, AAAI '99/IAAI '99, pages 439-446, Menlo Park, CA, USA. American Association for Artificial Intelligence.

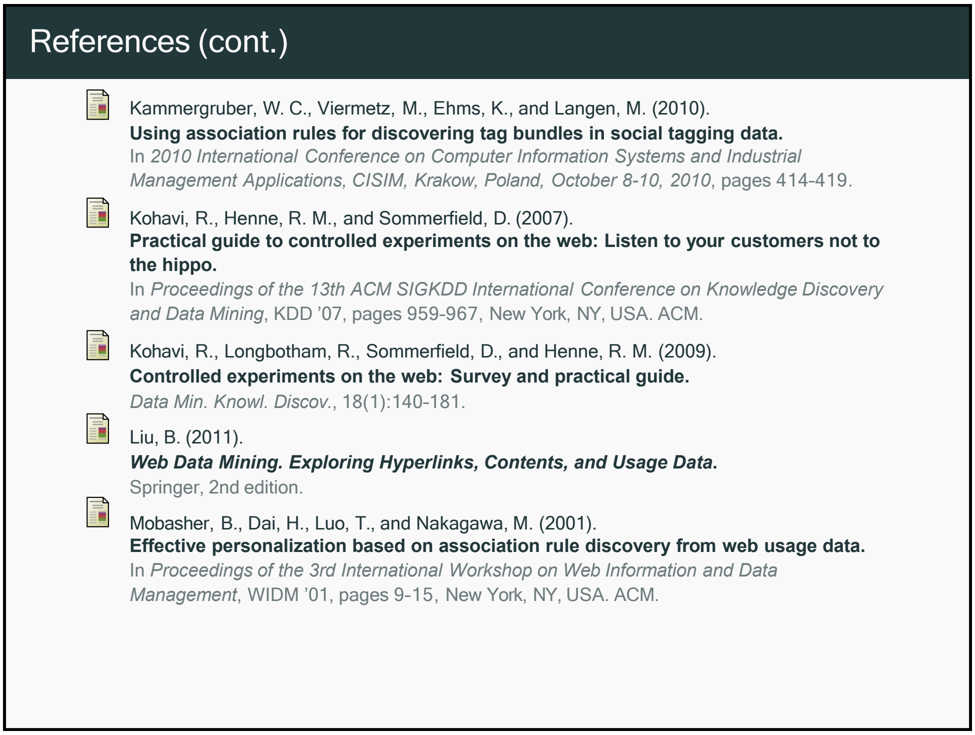
Han, J., Kamber, M., and Pei, J. (2011).
Data Mining: Concepts and Techniques.
Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition.

Huang, Z., Chen, H., and Zeng, D. (2004).
Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering.
ACM Trans. Inf. Syst., 22(1):116-142.

Jayanthi, S. (2013).
Web mining issues.
http://webminingissues.blogspot.pt/.

Jorge, A. (2016).
Web mining.
Slides.

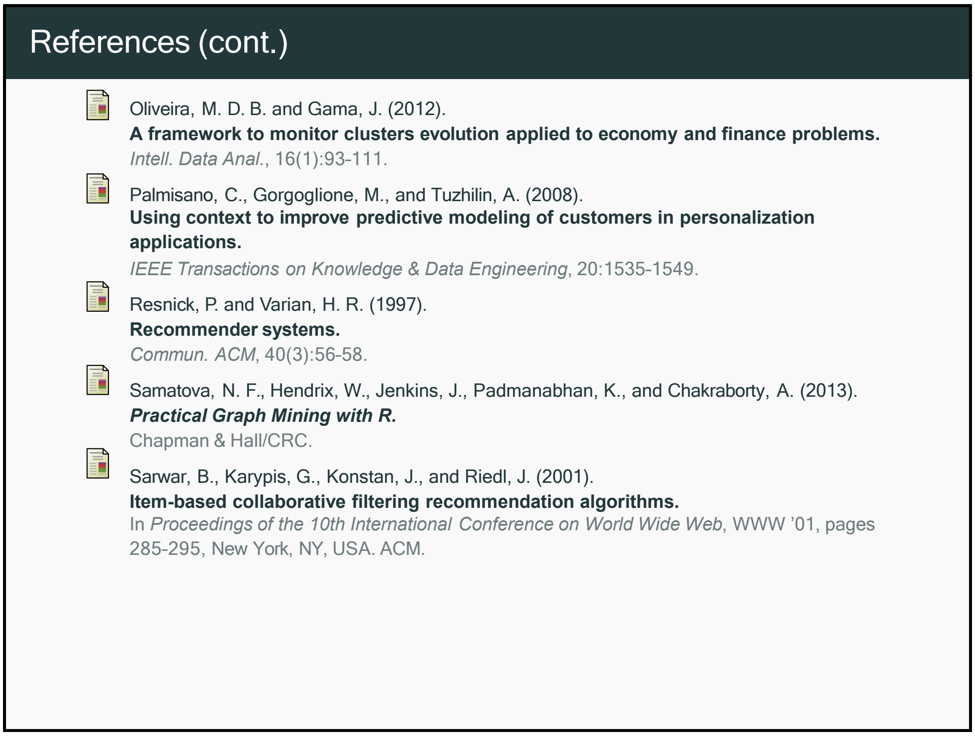
83



References (cont.)

- Kammergruber, W. C., Viermetz, M., Ehms, K., and Langen, M. (2010).
  Using association rules for discovering tag bundles in social tagging data.
  In *2010 International Conference on Computer Information Systems and Industrial Management Applications, CISIM, Krakow, Poland, October 8-10, 2010*, pages 414-419.
- Kohavi, R., Henne, R. M., and Sommerfield, D. (2007).
  Practical guide to controlled experiments on the web: Listen to your customers not to the hippo.
  In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’07, pages 959-967, New York, NY, USA. ACM.
- Kohavi, R., Longbotham, R., Sommerfield, D., and Henne, R. M. (2009).
  Controlled experiments on the web: Survey and practical guide.
  Data Min. Knowl. Discov., 18(1):140-181.
- Liu, B. (2011).
  Web Data Mining. Exploring Hyperlinks, Contents, and Usage Data.
  Springer, 2nd edition.
- Mobasher, B., Dai, H., Luo, T., and Nakagawa, M. (2001).
  Effective personalization based on association rule discovery from web usage data.
  In *Proceedings of the 3rd International Workshop on Web Information and Data Management*, WIDM ’01, pages 9-15, New York, NY, USA. ACM.

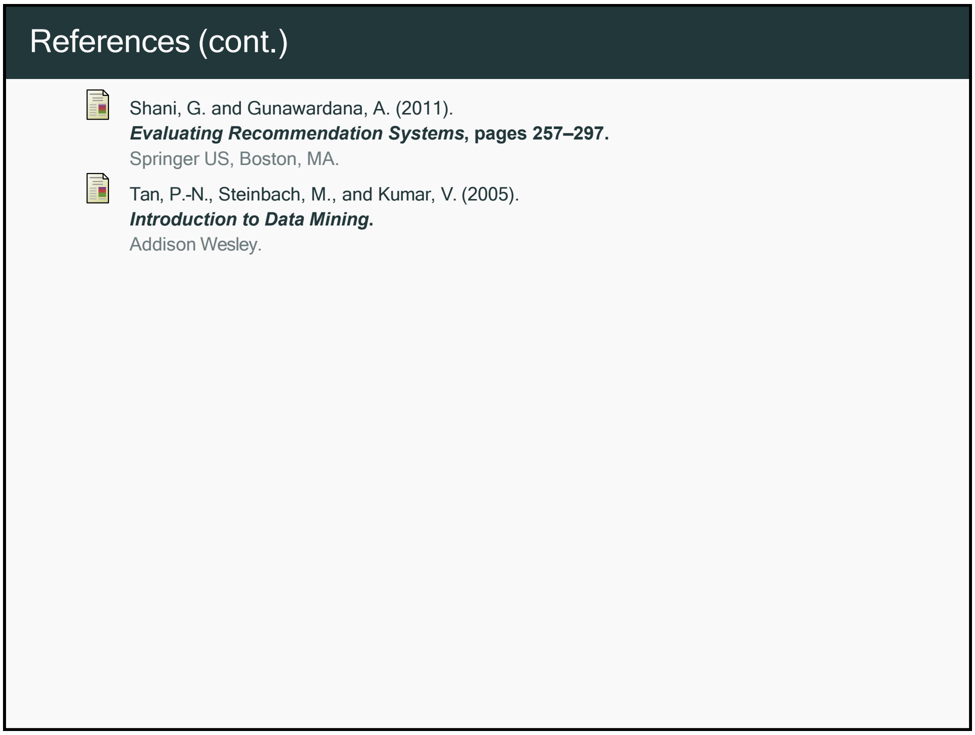
84



References (cont.)

- Oliveira, M. D. B. and Gama, J. (2012).
  A framework to monitor clusters evolution applied to economy and finance problems.
  Intell. Data Anal., 16(1):93-111.
- Palmisano, C., Gorgoglione, M., and Tuzhilin, A. (2008).
  Using context to improve predictive modeling of customers in personalization applications.
  IEEE Transactions on Knowledge & Data Engineering, 20:1535-1549.
- Resnick, P. and Varian, H. R. (1997).
  Recommender systems.
  Commun. ACM, 40(3):56-58.
- Samatova, N. F., Hendrix, W., Jenkins, J., Padmanabhan, K., and Chakraborty, A. (2013).
  Practical Graph Mining with R.
  Chapman & Hall/CRC.
- Sarwar, B., Karypis, G., Konstan, J., and Riedl, J. (2001).
  Item-based collaborative filtering recommendation algorithms.
  In *Proceedings of the 10th International Conference on World Wide Web*, WWW ’01, pages 285-295, New York, NY, USA. ACM.

85



References (cont.)

- Shani, G. and Gunawardana, A. (2011).
  ***Evaluating Recommendation Systems*, pages 257–297.**
  Springer US, Boston, MA.
- Tan, P.-N., Steinbach, M., and Kumar, V. (2005).
  ***Introduction to Data Mining*.**
  Addison Wesley.