UFMG UNIVERSIDADE FEDERAL DE MINAS GERAIS

Recommender Systems

Evaluation Metrics

Rodrygo L. T. Santos
rodrygo@dcc.ufmg.br

Why evaluate?

Gazillions of algorithms

◦ Collaborative, content-based, hybrid…

◦ Which one to choose?

Evaluation enables an informed choice

◦ Rigor of science

◦ Efficiency of practice

Recommender evaluation

Lessons from academia

- Evaluation methodologies
- User behavioral models
- Evaluation metrics

Lessons from industry

- What works in practice?

Evaluation metrics

Prediction accuracy

- *How well does it estimate absolute preferences?*

Decision support

- *How well does it return "good" things?*

Ranking accuracy

- *How well does it estimate relative preferences?*

Moving forward

Metrics tuned for specific purposes

- Sophisticated rank-based metrics
- Diversity, novelty, serendipity

Holistic evaluations

- Beyond just the recommendations
- Whole-page relevance

Evaluation metrics

General form: $\Delta(R_u, G_u)$

- R_u: items recommended to user u
- G_u: items relevant to user u

Metrics should be chosen according to the task

- Rating prediction, decision support, ranking

Accuracy metrics

Accuracy of a prediction

- Closeness to the actual preference

Actual preference unknown from system

- **Hidden** in an offline evaluation
- **Truly unknown** in an online evaluation

Typically measured by error metrics

Prediction accuracy

	i_1	i_2	i_3	i_4
u_1	5	3	3	?
u_2	5			3
u_3		1	2	1

$\hat{r}_{u_1 i_4} = 2.99$ vs. $r_{u_1 i_4} = 2$

how accurate is this prediction?

Prediction accuracy

	i_1	i_2	i_3	i_4
u_1	5	3	3	?
u_2	5			3
u_3		1	2	1

$\hat{r}_{u_1 i_4} = 2.99$ vs. $r_{u_1 i_4} = 2$

raw error

$$e_{u_1 i_4} = r_{u_1 i_4} - \hat{r}_{u_1 i_4}$$
$$= 2 - 2.99$$
$$= -0.99$$

Prediction accuracy

	i_1	i_2	i_3	i_4
u_1	5	3	3	?
u_2	5			3
u_3		1	2	1

$\hat{r}_{u_1 i_4} = 2.99$ vs. $r_{u_1 i_4} = 2$

absolute error

$$e_{u_1 i_4} = \left| r_{u_1 i_4} - \hat{r}_{u_1 i_4} \right|$$
$$= |2 - 2.99|$$
$$= 0.99$$

Prediction accuracy

	i_1	i_2	i_3	i_4
u_1	5	3	3	?
u_2	5			3
u_3		1	2	1

$\hat{r}_{u_1 i_4} = 2.99$ vs. $r_{u_1 i_4} = 2$

squared error

$$e_{u_1 i_4} = \left(r_{u_1 i_4} - \hat{r}_{u_1 i_4}\right)^2$$
$$= (2 - 2.99)^2$$
$$= 0.98$$

Prediction accuracy

Two wrongs don’t make a right!

- Absolute error removes direction

Large errors should be penalized more

- Squared error emphasizes discrepancies

Can’t assess effectiveness based on one prediction

- Average over multiple predictions

Prediction accuracy

	i_1	i_2	i_3	i_4
u_1	5	3	?	?
u_2	?			3
u_3		?	2	1

mean absolute error

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

Prediction accuracy

	i_1	i_2	i_3	i_4
u_1	5	3	?	?
u_2	?			3
u_3		?	2	1

mean squared error

$$\mathrm{MSE}(G) = \frac{1}{|G|}\sum_{(u,i)\in G}(r_{ui} - \hat{r}_{ui})^2$$

Prediction accuracy

	i_1	i_2	i_3	i_4
u_1	5	3	?	?
u_2	?			3
u_3		?	2	1

root mean squared error

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

Averaging errors

What could go wrong with average errors?

- We averaged over all ratings

What if a user has 10k ratings and another 10?

- The evaluation will be biased!

Alternative?

- Average over user averages

Averaging errors

Averaging over user averages

- $\mathrm{MAE}(G) = \frac{1}{|U|} \sum_{u \in U} \mathrm{MAE}(G_u)$
- $\mathrm{MSE}(G) = \frac{1}{|U|} \sum_{u \in U} \mathrm{MSE}(G_u)$
- $\mathrm{RMSE}(G) = \frac{1}{|U|} \sum_{u \in U} \mathrm{RMSE}(G_u)$

Accuracy metrics

Error metrics generally correlated

- MAE more interpretable, MSE more discriminative
- RMSE gives the best of both worlds

A few caveats

- Different rating scales are not comparable
- Errors can be dominated by popular users or items

Beyond accuracy

> *In industry, we care about keeping our users and making them happy, not improving accuracy of recommendations by 1%*

- Tao Ye, Senior Scientist at Amazon RecSys 2015, Industry Panel

Netflix Prize

COMPLETED

Home Rules Leaderboard Update

Leaderboard

Showing Test Score. Click here to show quiz score

Rank	Team Name	Best Test Score	% Improvement	Best Submit Time
Grand Prize - RMSE = 0.8567 - Winning Team: BellKor's Pragmatic Chaos				
1	BellKor's Pragmatic Chaos	0.8567	10.06	2009-07-26 18:18:28
2	The Ensemble	0.8567	10.06	2009-07-26 18:38:22
3	Grand Prize Team	0.8582	9.90	2009-07-10 21:24:40
4	Opera Solutions and Vandelay United	0.8588	9.84	2009-07-10 01:12:31
5	Vandelay Industries !	0.8591	9.81	2009-07-10 00:32:20
6	PragmaticTheory	0.8594	9.77	2009-06-24 12:06:56
7	BellKor in BigChaos	0.8601	9.70	2009-05-13 08:14:09

Beyond accuracy

Decision support is key

- Does the recommender return “good” items?

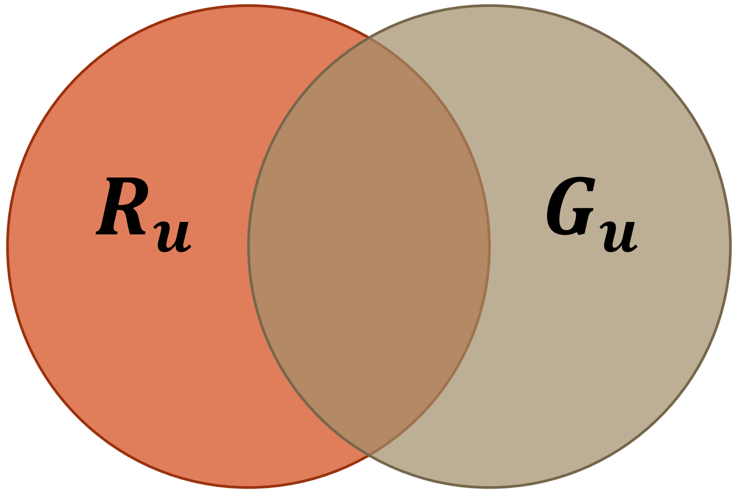
Say 1-3 stars is bad, 4-5 stars is good

- Recommender #1 says 3 stars, user says 1 star
- Recommender #2 says 4 stars, user says 2 stars

Recommender #2 **misleads** the user

Precision and recall

Given a user $\boldsymbol{u}$



$\boldsymbol{R}_u$: recommended items
$\boldsymbol{G_u}$: relevant items

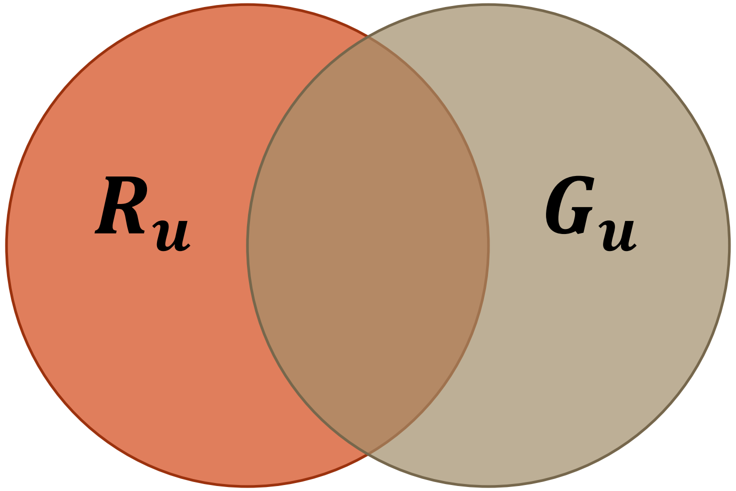
Precision

- Percentage of recommended items that are relevant

$$\mathrm{Prec}(R_u, G_u) = \frac{|R_u \cap G_u|}{|R_u|}$$

Precision and recall

Given a user $\boldsymbol{u}$



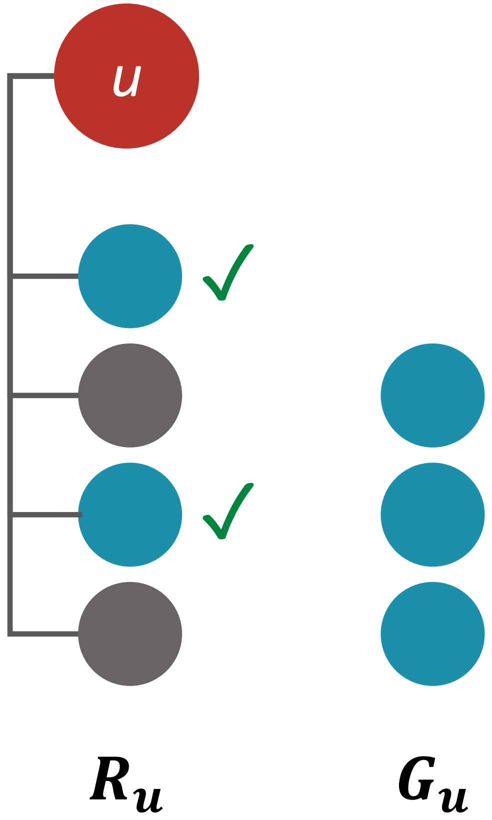
$\boldsymbol{R}_u$: recommended items
$\boldsymbol{G}_{\boldsymbol{u}}$: relevant items

Recall

- Percentage of relevant items that are recommended

$$\mathrm{Rec}(R_u, G_u) = \frac{|R_u \cap G_u|}{|G_u|}$$

Precision and recall



Precision

- $\mathrm{Prec}(R_u, G_u) = \frac{|R_u \cap G_u|}{|R_u|} = \frac{2}{4} = 0.50$

Recall

- $\mathrm{Rec}(R_u, G_u) = \frac{|R_u \cap G_u|}{|G_u|} = \frac{2}{3} = 0.67$

Precision and recall

Precision is about having mostly useful stuff in a recommendation

- Not wasting the user's time

Key assumption

- There is more useful stuff than you want to examine

Recall is about not missing useful stuff in a recommendation

- Not making a bad oversight

Key assumption

- You have time to filter through recommendations

We can also combine both

$$\mathrm{F1}(R,G) = \frac{2\,\mathrm{Prec}(R,G)\,\mathrm{Rec}(R,G)}{\mathrm{Prec}(R,G) + \mathrm{Rec}(R,G)}$$

Beyond decision support

Modern item catalogs are huge

- User may not be willing to inspect large sets

Consider top-5 rankings

- Recommender #1: + + + + –
- Recommender #2: – + + + +

Recommender #2 **misplaces** a highly visible item

Summarizing a ranking

Calculating recall and precision at fixed rank positions

- e.g., Prec@10, Rec@10

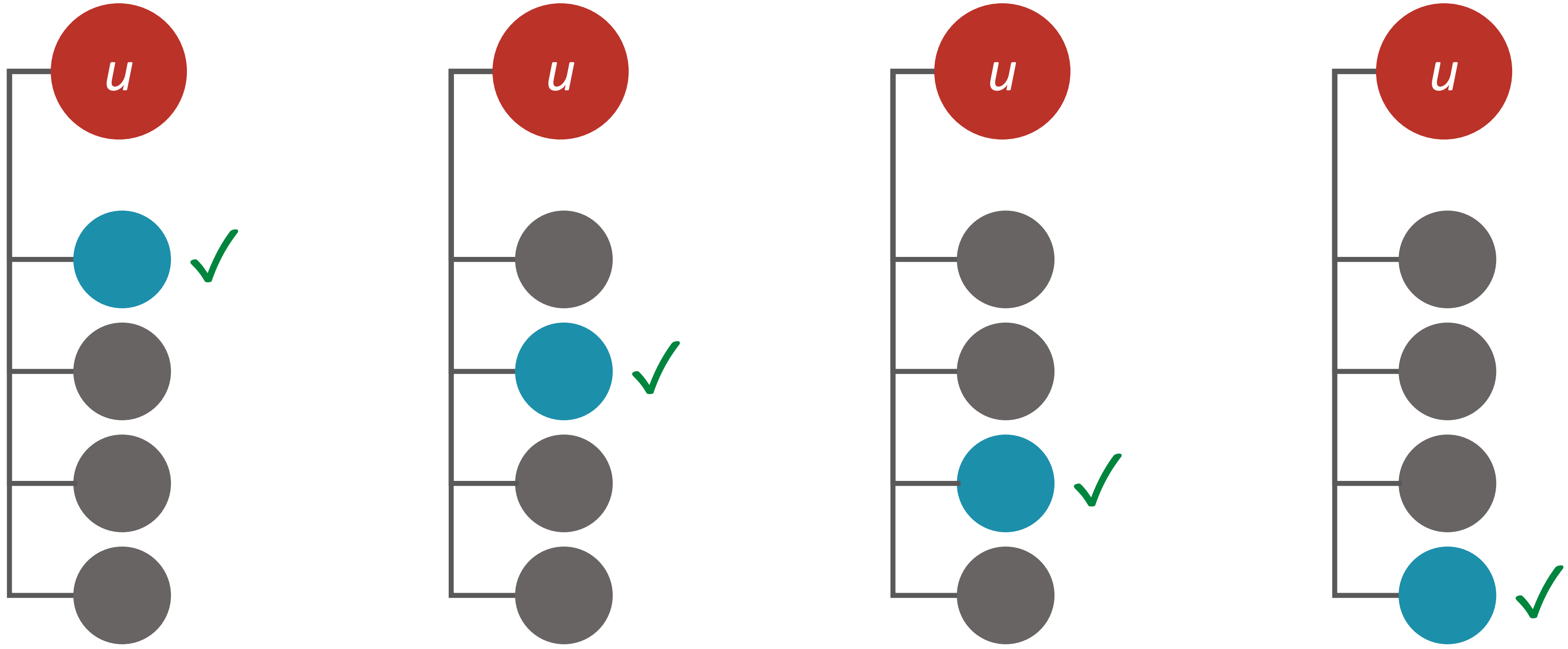
Calculating precision at standard recall levels

- e.g., Prec@Rec=30%

Problem

- Set-based metrics are blind within the set

Position blindness



These have exactly the same Prec@4 (0.25)

- Are they equally good?

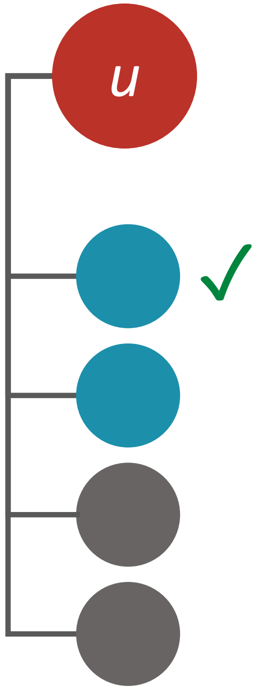
Ranking metrics

Why ranking?

- Place items in order of preference

Key assumption

- Users will inspect recommended items from top to bottom (or left to right)

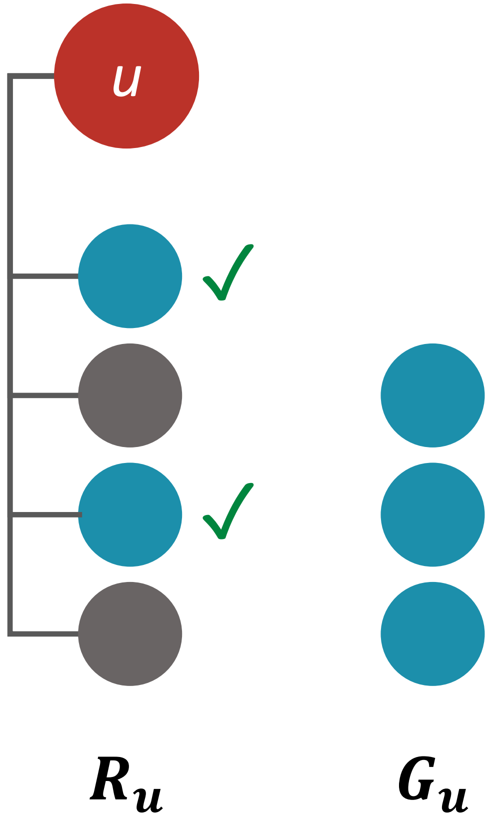


Average precision (AP)

Simple idea: averaging precision values at the ranking positions where relevant items were found

- $\mathrm{AP}(R, G) = \frac{1}{|G|} \sum_{c=1}^{k} 1\left(r_{ui_c} > 0\right) \mathrm{Prec@}c$

Average precision (AP)



Average precision

- $\mathrm{AP}(R, G) = \frac{1}{|G|} \sum_{c=1}^{k} 1\left(r_{ui_c} > 0\right) \mathrm{Prec@}c$

$$= \frac{1}{3}(\mathrm{Prec@}1 + \mathrm{Prec@}3)$$

$$= \frac{1}{3}\left(\frac{1}{1} + \frac{2}{3}\right)$$

$$= \frac{5}{9} = 0.55$$

Average precision (AP)

Simple idea: averaging precision values at the ranking positions where relevant items were found

- $\mathrm{AP}(R, G) = \frac{1}{|G|} \sum_{c=1}^{k} 1\left(r_{ui_c} > 0\right) \mathrm{Prec@}c$

In practice, take the mean (MAP) across users

- $\mathrm{MAP} = \frac{1}{|U|} \sum_{u \in U} \mathrm{AP}(R_u, G_u)$

Reciprocal rank (RR)

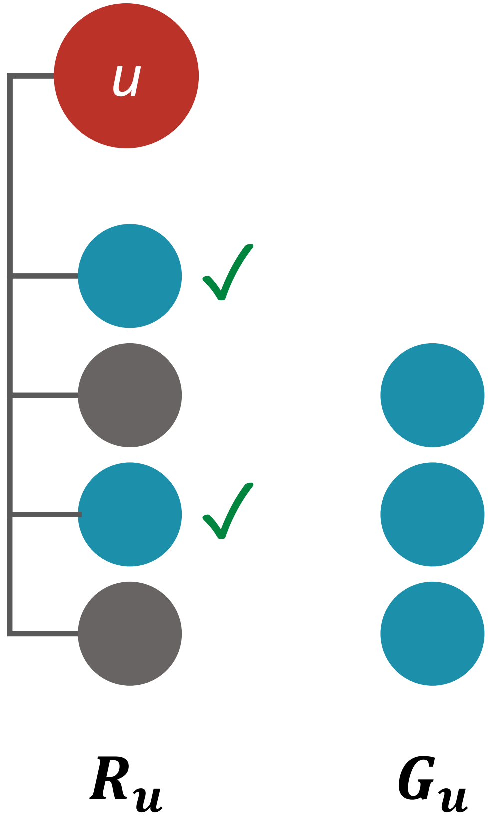
Measures how deep the user has to dig in the recommended items to find the first relevant item

- $\mathrm{RR}(R, G) = 1/c$ (c: position of the first relevant)

Mean reciprocal rank averages across users

- $\mathrm{MRR} = \frac{1}{|U|} \sum_{u \in U} \mathrm{RR}(R_u, G_u)$

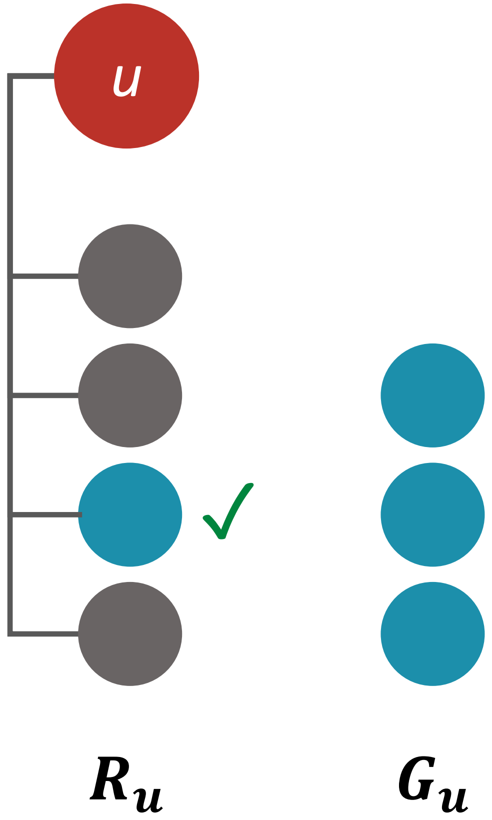
Reciprocal rank (RR)



Reciprocal rank

- $\mathrm{RR}(R, G) = \frac{1}{c}$
  $= \frac{1}{1} = 1.00$

Reciprocal rank (RR)



Reciprocal rank

- $\mathrm{RR}(R, G) = \dfrac{1}{c}$

$$= \frac{1}{3} = 0.33$$

Discounted cumulative gain (DCG)

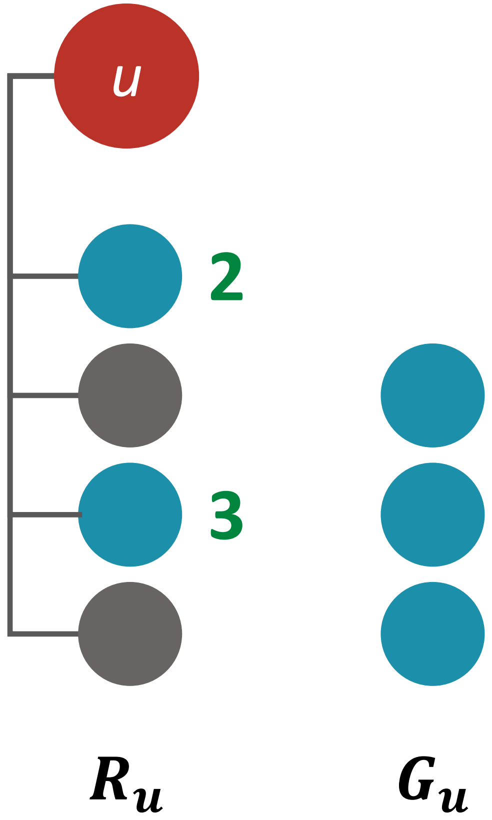
Measure utility of item at each position

- $\mathrm{DCG}(R, G) = \sum_{c=1}^{k} \frac{r_{ui_c}}{\log_2(c+1)}$

linear gain (e.g., in a graded scale)

position-based discount

Discounted cumulative gain (DCG)



Discounted cumulative gain

- $\mathrm{DCG}(R, G) = \sum_{c=1}^{k} \frac{r_{ui_c}}{\log_2(c+1)}$

$$= \frac{2}{\log_2 1+1} + \frac{3}{\log_2 3+1}$$

$$= \frac{2}{1} + \frac{3}{2}$$

$$= 2 + 1.5$$

$$= 3.5$$

Discounted cumulative gain (DCG)

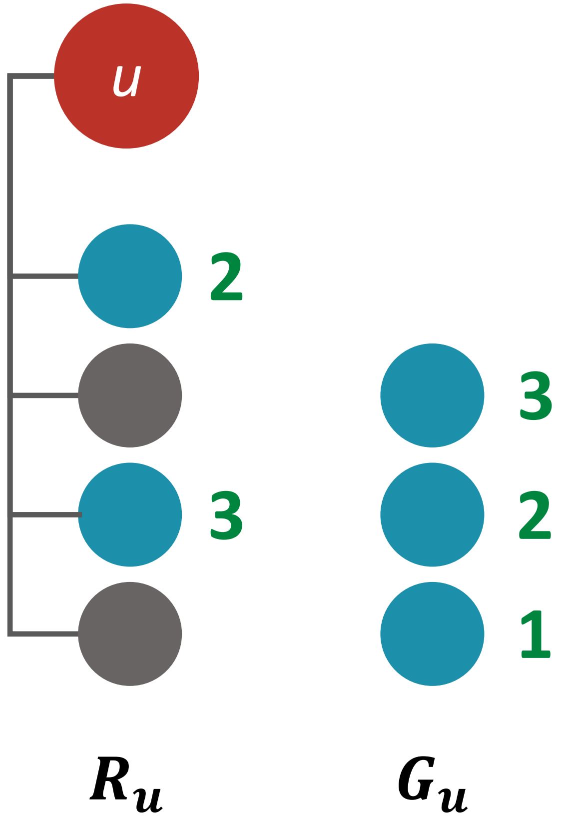
Measure utility of item at each position

- $\mathrm{DCG}(R, G) = \sum_{c=1}^{k} \frac{r_{ui_c}}{\log_2(c+1)}$ linear gain (e.g., in a graded scale) position-based discount

Could also emphasize larger gains

- $\mathrm{DCG}(R, G) = \sum_{c=1}^{k} \frac{2^{r_{ui_c}}-1}{\log_2(c+1)}$ exponential gain position-based discount

Ideal discounted cumulative gain (iDCG)



Ideal discounted cumulative gain

- $\mathrm{iDCG}(R, G) = \sum_{c=1}^{k} \frac{r_{ui_c}}{\log_2(c+1)}$

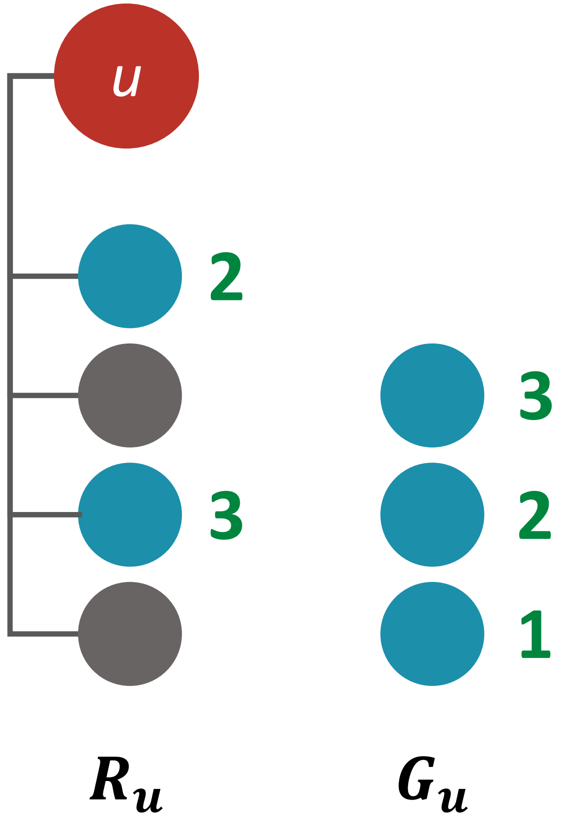
$$= \frac{3}{\log_2 1+1} + \frac{2}{\log_2 2+1} + \frac{1}{\log_2 3+1}$$

$$= \frac{3}{1} + \frac{2}{1.58} + \frac{3}{2}$$

$$= 3 + 1.26 + 0.5$$

$$= 4.76$$

Norm. discounted cumulative gain (nDCG)



Normalized discounted cumulative gain

- $\text{nDCG}(R, G) = \frac{\text{DCG}(R,G)}{\text{iDCG}(R,G)}$

$$= \frac{3.5}{4.76}$$

$$= 0.74$$

Average nDCG

In practice, nDCG is averaged across all test users

- $\mathrm{nDCG} = \frac{1}{|U|} \sum_{u \in U} \frac{\mathrm{DCG}(R_u, G_u)}{\mathrm{iDCG}(R_u, G_u)}$

Ranking-based metrics

Several metrics to measure a recommender’s ability to order the recommended items

- Mostly borrowed from search evaluation

nDCG increasingly common

- MRR also used

Business metrics

We are interested in satisfying the user

- Accuracy metrics
- Decision support metrics
- Ranking metrics

But also the recommendation provider

- Coverage, diversity, serendipity

Coverage

Measures the percentage of products for which a recommender can make a prediction

- Or a prediction that’s personalized
- Or a prediction above a confidence threshold
  - e.g., how many 5-stars movies can we recommend?

Business interest: reach the entire catalog

Diversity

Measures how different the recommendations are

- With respect to other recommended items, items the user knows of, items everyone knows of
- e.g., intra-list diversity (ILD) is the average pairwise dissimilarity among recommended item

Business interest: sales diversity

Serendipity

Measures “the occurrence of events by chance in a happy or beneficial way”

- In RS: surprising, delightful unexpectedness

Several ways to operationalize

- Typically, based on rarity

Summary

Several metrics for different purposes

- No one-size-fits-all solution
- Different metrics, different quality estimates

Metrics may not well correlate with practice

- Must look outside the box

References

Recommender Systems: An Introduction (Ch. 7)

Recommender Systems Handbook (Ch. 8)

Recommender Systems: The Textbook (Ch. 7)

Statistical Methods for Recommender Systems (Ch. 4)