New approaches for evaluation: correctness and freshness

Pablo Sánchez Rus M. Mesas Alejandro Bellogín

Universidad Autónoma de Madrid
Escuela Politécnica Superior
Departamento de Ingeniería Informática

V Congreso Español de
Recuperación de Información (CERI 2018)

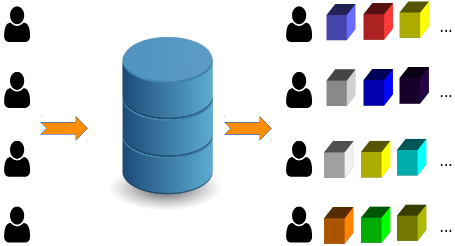
Outline

1. Recommender Systems
2. Freshness
3. Correctness
4. Experiments
5. Conclusions and future work

Outline

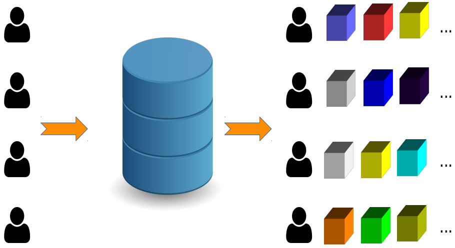
1. Recommender Systems
2. Freshness
3. Correctness
4. Experiments
5. Conclusions and future work

Recommender Systems



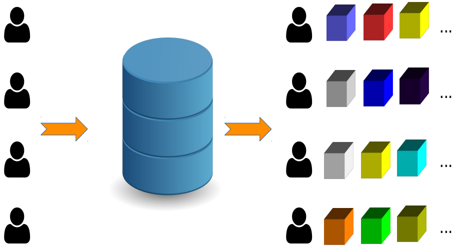
- Suggest **new items** to users based on their tastes and needs

Recommender Systems



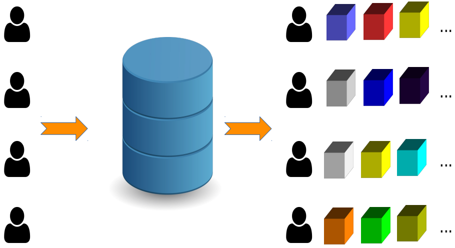
- Suggest **new items** to users based on their tastes and needs
- Measure the **quality** of recommendations. How?

Recommender Systems



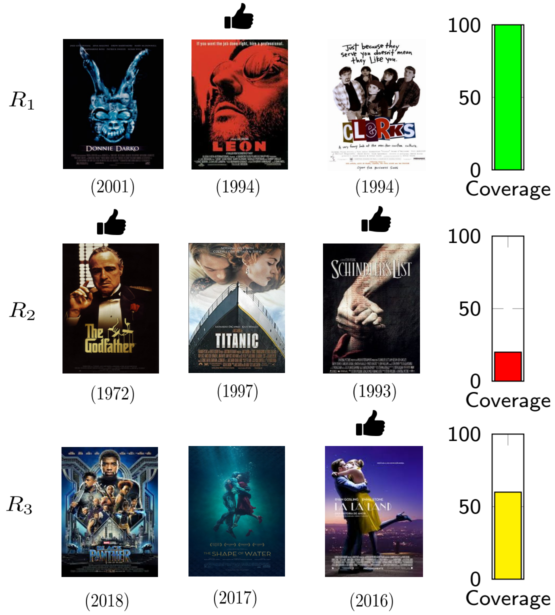
- Suggest **new items** to users based on their tastes and needs
- Measure the **quality** of recommendations. How?
- Several evaluation dimensions:
  Error, Ranking, Novelty / Diversity

Recommender Systems

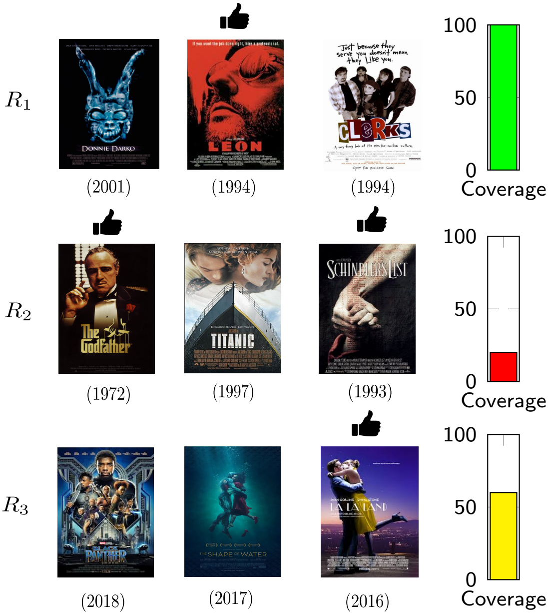


- Suggest **new items** to users based on their tastes and needs
- Measure the **quality** of recommendations. How?
- Several evaluation dimensions:
  Error, Ranking, Novelty / Diversity
- We will focus on **Freshness** and **Correctness** (from **Sánchez and Bellogín (2018); Mesas and Bellogín (2017)**)

Different notions of quality



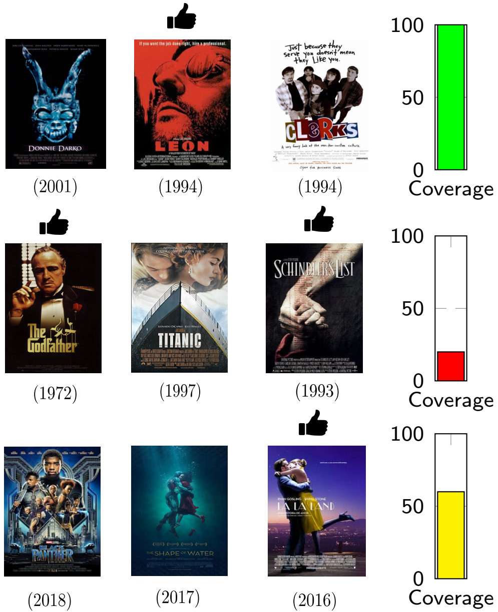
Different notions of quality



- Best in Relevance?
  - $R_2 > R_1 > R_3$

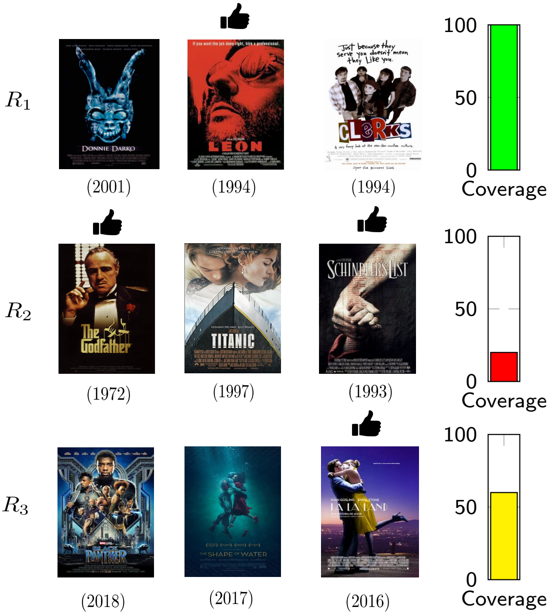
Different notions of quality

R_1, R_2, R_3



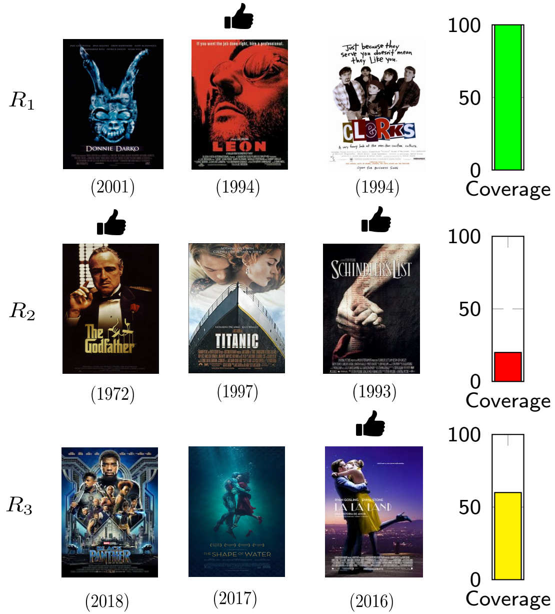
- Best in Relevance?
  - $R_2 > R_1 > R_3$
- Best in Novelty?
  - $R_1 > R_3 > R_2$

Different notions of quality



- Best in Relevance?
  - $R_2 > R_1 > R_3$
- Best in Novelty?
  - $R_1 > R_3 > R_2$
- Best in **Freshness**?
  - $R_3 > R_1 > R_2$

Different notions of quality



- Best in Relevance?
  - $R_2 > R_1 > R_3$
- Best in Novelty?
  - $R_1 > R_3 > R_2$
- Best in **Freshness**?
  - $R_3 > R_1 > R_2$
- Best in **Cov-Rel Tradeoff**?
  - $R_1 > R_3 > R_2$??
  - $R_1 > R_2 > R_3$??

Outline

1 Recommender Systems

2 Freshness

3 Correctness

4 Experiments

5 Conclusions and future work

Preliminaries

- Framework proposed in **Vargas and Castells (2011)**

$$m(R_u \mid \theta) = C \sum_{i_n \in R_u} \mathrm{disc}(n) p(rel \mid i_n, u) \mathrm{nov}(i_n \mid \theta) \tag{1}$$

Preliminaries

- Framework proposed in **Vargas and Castells (2011)**

$$m(R_u \mid \theta) = C \sum_{i_n \in R_u} \mathrm{disc}(n) p(rel \mid i_n, u) \mathrm{nov}(i_n \mid \theta) \tag{1}$$

- Where:
  - R_u items recommended to user u
  - θ contextual variable (e.g., the user profile)
  - $\mathrm{disc}(n)$ is a discount model (e.g. NDCG)
  - $p(rel \mid i_n, u)$ relevance component
  - $\mathrm{nov}(i_n \mid \theta)$ novelty model

Preliminaries

- Framework proposed in **Vargas and Castells (2011)**

$$m(R_u \mid \theta) = C \sum_{i_n \in R_u} \mathrm{disc}(n) p(rel \mid i_n, u) \mathrm{nov}(i_n \mid \theta) \tag{1}$$

- With this framework we can derive multiple metrics, however, all of them are *time-agnostic*

Preliminaries

- Framework proposed in **Vargas and Castells (2011)**

$$m(R_u \mid \theta_t) = C \sum_{i_n \in R_u} \mathrm{disc}(n) p(rel \mid i_n, u) \boxed{\mathrm{nov}(i_n \mid \theta_t)} \tag{1}$$

- With this framework we can derive multiple metrics, however, all of them are *time-agnostic*
- We propose to replace the novelty component defining new **time-aware novelty models**

Time-Aware Novelty Metrics

- Classic metrics do not provide any information about the evolution of the items: we can recommend relevant but well-known (old) items

Time-Aware Novelty Metrics

- Classic metrics do not provide any information about the evolution of the items: we can recommend relevant but well-known (old) items
- Every item in the system can be modeled with a temporal representation:

$$\theta_t = \{\theta_t(i)\} = \{(i, \langle t_1(i), \cdots, t_n(i)\rangle)\} \tag{2}$$

Time-Aware Novelty Metrics

- Classic metrics do not provide any information about the evolution of the items: we can recommend relevant but well-known (old) items
- Every item in the system can be modeled with a temporal representation:

$$\theta_t = \{\theta_t(i)\} = \{(i, \langle t_1(i), \cdots, t_n(i)\rangle)\} \quad (2)$$

- Two different sources for the timestamps:

Time-Aware Novelty Metrics

- Classic metrics do not provide any information about the evolution of the items: we can recommend relevant but well-known (old) items
- Every item in the system can be modeled with a temporal representation:

$$\theta_t = \{\theta_t(i)\} = \{(i, \langle t_1(i), \cdots, t_n(i)\rangle)\} \tag{2}$$

- Two different sources for the timestamps:
  - Metadata information: release date (movies or songs), creation time, etc.

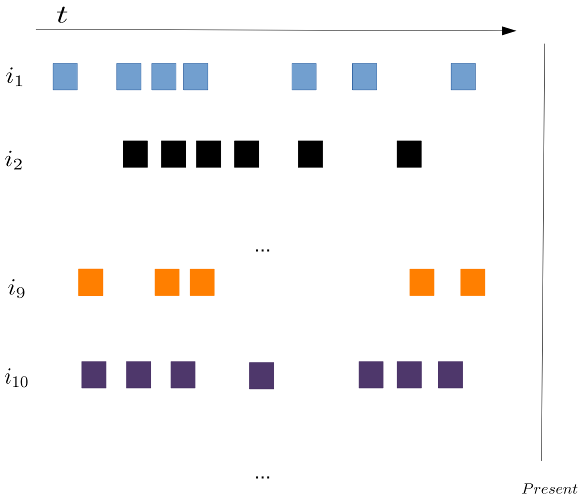
Time-Aware Novelty Metrics

- Classic metrics do not provide any information about the evolution of the items: we can recommend relevant but well-known (old) items
- Every item in the system can be modeled with a temporal representation:

$$\theta_t = \{\theta_t(i)\} = \{(i, \langle t_1(i), \cdots, t_n(i)\rangle)\} \tag{2}$$

- Two different sources for the timestamps:
  - Metadata information: release date (movies or songs), creation time, etc.
  - Rating history of the items

Time-Aware Novelty Metrics



Modeling time profiles for items

- How can we aggregate the temporal representation?

Modeling time profiles for items

- How can we aggregate the temporal representation?
- We explored four possibilities:

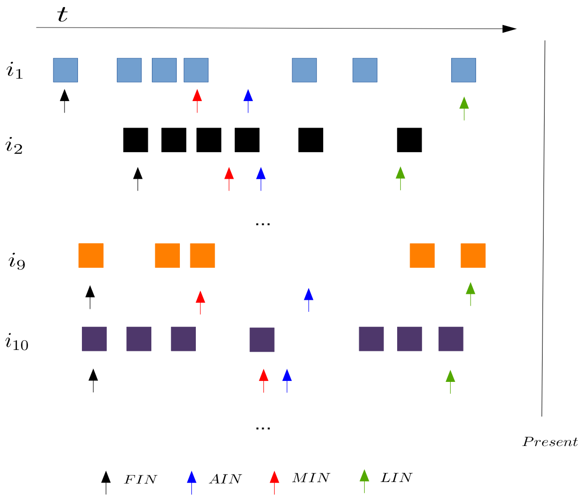
Modeling time profiles for items

- How can we aggregate the temporal representation?
- We explored four possibilities:
  - Take the first interaction (FIN)
  - Take the last interaction (LIN)
  - Take the average of the ratings times (AIN)
  - Take the median of the ratings times (MIN)

Modeling time profiles for items

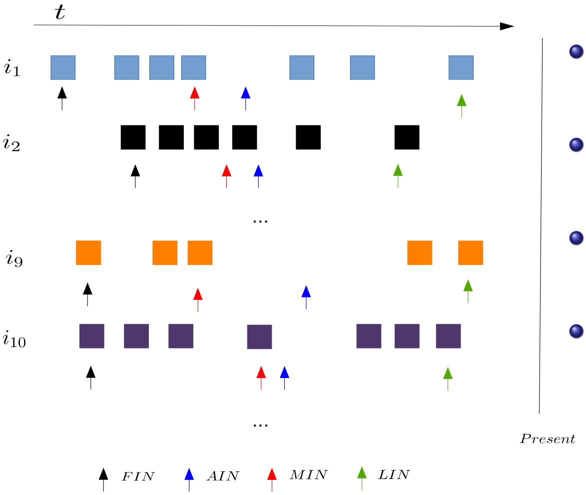
- How can we aggregate the temporal representation?
- We explored four possibilities:
  - Take the first interaction (FIN)
  - Take the last interaction (LIN)
  - Take the average of the ratings times (AIN)
  - Take the median of the ratings times (MIN)
- Each case defines a function $f(\theta_t(i))$

Modeling time profiles for items: an example



Modeling time profiles for items: an example

- Which model represents better the freshness of the items?



- FIN?
  - $i_2 > i_{10} > i_9 > i_1$
- LIN?
  - $i_9 > i_1 > i_{10} > i_2$
- MIN?
  - $i_{10} > i_2 > i_9 > i_1$
- AIN?
  - $i_9 > i_{10} > i_2 > i_1$

Outline

1. Recommender Systems
2. Freshness
3. Correctness
4. Experiments
5. Conclusions and future work

Motivation

- Goal: balancing coverage and precision

Motivation

- Goal: balancing coverage and precision
- Some researchers (**Herlocker et al. (2004) Gunawardana and Shani (2015)**) alerted this is still an open problem in Recommender Systems evaluation

Motivation

- Goal: balancing coverage and precision
- Some researchers (**Herlocker et al. (2004) Gunawardana and Shani (2015)**) alerted this is still an open problem in Recommender Systems evaluation
- Typical situation: recommendations with low confidence should not be presented to the user (coverage is reduced at the expense of (potentially) more relevant recommendations)

Our proposal: Correctness metrics

- Adapted from Question Answering (**Peñas and Rodrigo (2011)**)

Our proposal: Correctness metrics

- Adapted from Question Answering (**Peñas and Rodrigo (2011)**)
- Each question has several options but only one answer is correct

Our proposal: Correctness metrics

- Adapted from Question Answering (**Peñas and Rodrigo (2011)**)
- Each question has several options but only one answer is correct
- If an answer is not given, it should not be considered as incorrect (the algorithm *decided not to recommend*)

Our proposal: Correctness metrics

- Adapted from Question Answering (**Peñas and Rodrigo (2011)**)
- Each question has several options but only one answer is correct
- If an answer is not given, it should not be considered as incorrect (the algorithm *decided not to recommend*)
- Applied to recommenders: if two systems have the same number of relevant items but one has retrieved less items, it should be better than the other one

Our proposal: Correctness metrics

- Based on users:

$$\text{User Correctness} = \frac{1}{N}\left(TP(u) + TP(u)\frac{NR(u)}{N}\right) \tag{3}$$

$$\text{Recall User Correctness} = \frac{1}{N}\left(TP(u) + \frac{TP(u)}{|T(u)|}NR(u)\right) \tag{4}$$

Our proposal: Correctness metrics

- Based on users:

$$\text{User Correctness} = \frac{1}{N}\left(TP(u) + TP(u)\frac{NR(u)}{N}\right) \quad (3)$$

$$\text{Recall User Correctness} = \frac{1}{N}\left(TP(u) + \frac{TP(u)}{|T(u)|}NR(u)\right) \quad (4)$$

- where
  - $TP(u)$: number of relevant items that we are recommending to the user
  - $FP(u)$: number of non-relevant items that we are recommending to the user
  - N: cutoff
  - $NR(u) : N - (TP + FP)$
  - $|T(u)|$: number of relevant items in the test of user u

Experiments

1. Recommender Systems
2. Freshness
3. Correctness
4. Experiments
5. Conclusions and future work

Freshness results

- Are the recommendations obtained by different algorithms temporally novel (fresh)?
- Do the different novelty models produce similar results?

Freshness results: MovieLens (temporal split)

Algorithm	P	NDCG	USC	No relevance FIN	LIN	AIN	MIN
Rnd	0.0009	0.0010	**100.0**	0.5573†	0.9834	0.6993†	0.6711†
IdAsc	0.0099	0.0162	100.0‡	0.0716	0.9991	0.3550	0.2437
IdDec	0.0000	0.0000	100.0†	**0.9995**	0.9995	**0.9995**	**0.9995**
Pop	**0.1027**	**0.1110**	100.0	0.0781	0.9999‡	0.4361	0.3772
UB	0.0498‡	0.0618‡	17.8	0.2431	0.9999†	0.5835	0.5594
TD	0.0420	0.0520	17.8	0.6108‡	**0.9999**	0.7838‡	0.7710‡
HKV	0.0498†	0.0611†	17.8	0.3068	0.9998	0.6122	0.5885

Freshness results: MovieLens (temporal split)

Algorithm	P	NDCG	USC	No relevance FIN	LIN	AIN	MIN
Rnd	0.0009	0.0010	**100.0**	0.5573†	0.9834	0.6993†	0.6711†
IdAsc	0.0099	0.0162	100.0‡	0.0716	0.9991	0.3550	0.2437
IdDec	0.0000	0.0000	100.0†	**0.9995**	0.9995	**0.9995**	**0.9995**
Pop	**0.1027**	**0.1110**	100.0	0.0781	0.9999‡	0.4361	0.3772
UB	0.0498‡	0.0618‡	17.8	0.2431	0.9999†	0.5835	0.5594
TD	0.0420	0.0520	17.8	0.6108‡	**0.9999**	0.7838‡	0.7710‡
HKV	0.0498†	0.0611†	17.8	0.3068	0.9998	0.6122	0.5885

- Relevance metrics (P and NDCG), User Coverage (USC) and Freshness without relevance component (FIN, LIN, AIN, MIN)

Freshness results: MovieLens (temporal split)

Algorithm	P	NDCG	USC	No relevance FIN	LIN	AIN	MIN
Rnd	0.0009	0.0010	**100.0**	0.5573†	0.9834	0.6993†	0.6711†
IdAsc	0.0099	0.0162	100.0‡	0.0716	0.9991	0.3550	0.2437
IdDec	0.0000	0.0000	100.0†	**0.9995**	0.9995	**0.9995**	**0.9995**
Pop	**0.1027**	**0.1110**	100.0	0.0781	0.9999‡	0.4361	0.3772
UB	0.0498‡	0.0618‡	17.8	0.2431	0.9999†	0.5835	0.5594
TD	0.0420	0.0520	17.8	0.6108‡	**0.9999**	0.7838‡	0.7710‡
HKV	0.0498†	0.0611†	17.8	0.3068	0.9998	0.6122	0.5885

- Relevance metrics (P and NDCG), User Coverage (USC) and Freshness without relevance component (FIN, LIN, AIN, MIN)
- Very low coverage for personalized recommenders (due to temporal split)

Freshness results: MovieLens (temporal split)

Algorithm	P	NDCG	USC	No relevance FIN	LIN	AIN	MIN
Rnd	0.0009	0.0010	**100.0**	0.5573†	0.9834	0.6993†	0.6711†
IdAsc	0.0099	0.0162	100.0‡	0.0716	0.9991	0.3550	0.2437
IdDec	0.0000	0.0000	100.0†	**0.9995**	0.9995	**0.9995**	**0.9995**
Pop	**0.1027**	**0.1110**	100.0	0.0781	0.9999‡	0.4361	0.3772
UB	0.0498‡	0.0618‡	17.8	0.2431	0.9999†	0.5835	0.5594
TD	0.0420	0.0520	17.8	0.6108‡	**0.9999**	0.7838‡	0.7710‡
HKV	0.0498†	0.0611†	17.8	0.3068	0.9998	0.6122	0.5885

- Relevance metrics (P and NDCG), User Coverage (USC) and Freshness without relevance component (FIN, LIN, AIN, MIN)
- Very low coverage for personalized recommenders (due to temporal split)
- Data bias: the higher the id, the fresher the item (and the lower the id, the older the item)

Freshness results: MovieLens (temporal split)

Algorithm	P	NDCG	USC	No relevance FIN	LIN	AIN	MIN
Rnd	0.0009	0.0010	**100.0**	0.5573†	0.9834	0.6993†	0.6711†
IdAsc	0.0099	0.0162	100.0‡	0.0716	0.9991	0.3550	0.2437
IdDec	0.0000	0.0000	100.0†	**0.9995**	0.9995	**0.9995**	**0.9995**
Pop	**0.1027**	**0.1110**	100.0	0.0781	0.9999‡	0.4361	0.3772
UB	0.0498‡	0.0618‡	17.8	0.2431	0.9999†	0.5835	0.5594
TD	0.0420	0.0520	17.8	0.6108‡	**0.9999**	0.7838‡	0.7710‡
HKV	0.0498†	0.0611†	17.8	0.3068	0.9998	0.6122	0.5885

- Relevance metrics (P and NDCG), User Coverage (USC) and Freshness without relevance component (FIN, LIN, AIN, MIN)
- Very low coverage for personalized recommenders (due to temporal split)
- Data bias: the higher the id, the fresher the item (and the lower the id, the older the item)
- Popularity bias

Freshness results: Popularity bias

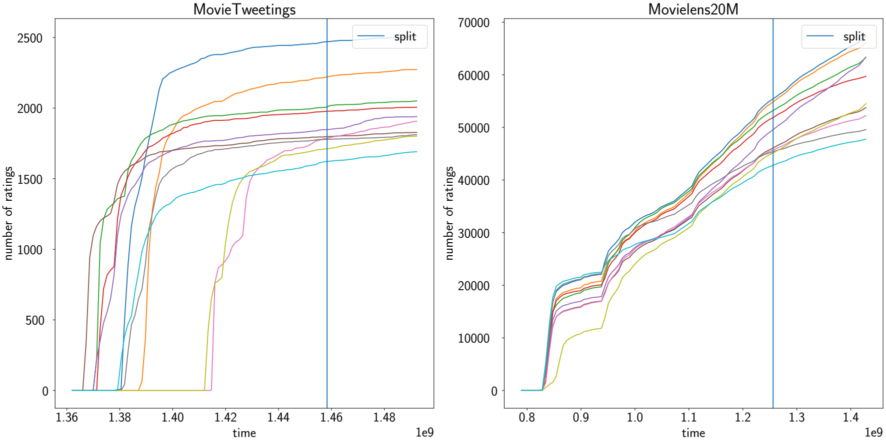


Figure: Top 10 most popular items in the training set of each dataset: MovieTweetings (left) and MovieLens (right).

Freshness results: MovieLens (temporal split)

Algorithm	P	NDCG	USC	No relevance FIN	LIN	AIN	MIN
Rnd	0.0009	0.0010	**100.0**	0.5573†	0.9834	0.6993†	0.6711†
IdAsc	0.0099	0.0162	100.0‡	0.0716	0.9991	0.3550	0.2437
IdDec	0.0000	0.0000	100.0†	**0.9995**	0.9995	**0.9995**	**0.9995**
Pop	**0.1027**	**0.1110**	100.0	0.0781	0.9999‡	0.4361	0.3772
UB	0.0498‡	0.0618‡	17.8	0.2431	0.9999†	0.5835	0.5594
TD	0.0420	0.0520	17.8	0.6108‡	**0.9999**	0.7838‡	0.7710‡
HKV	0.0498†	0.0611†	17.8	0.3068	0.9998	0.6122	0.5885

- Temporal recommenders less competitive in this dataset (no completely realistic timestamps)

Freshness results: MovieLens (temporal split)

Algorithm	P	NDCG	USC	No relevance FIN	LIN	AIN	MIN
Rnd	0.0009	0.0010	**100.0**	0.5573†	0.9834	0.6993†	0.6711†
IdAsc	0.0099	0.0162	100.0‡	0.0716	0.9991	0.3550	0.2437
IdDec	0.0000	0.0000	100.0†	**0.9995**	0.9995	**0.9995**	**0.9995**
Pop	**0.1027**	**0.1110**	100.0	0.0781	0.9999‡	0.4361	0.3772
UB	0.0498‡	0.0618‡	17.8	0.2431	0.9999†	0.5835	0.5594
TD	0.0420	0.0520	17.8	0.6108‡	**0.9999**	0.7838‡	0.7710‡
HKV	0.0498†	0.0611†	17.8	0.3068	0.9998	0.6122	0.5885

- Temporal recommenders less competitive in this dataset (no completely realistic timestamps)
- LIN not very useful

Freshness results: MovieLens (temporal split)

Algorithm	P	NDCG	USC	No relevance FIN	LIN	AIN	MIN
Rnd	0.0009	0.0010	**100.0**	0.5573†	0.9834	0.6993†	0.6711†
IdAsc	0.0099	0.0162	100.0‡	0.0716	0.9991	0.3550	0.2437
IdDec	0.0000	0.0000	100.0†	**0.9995**	0.9995	**0.9995**	**0.9995**
Pop	**0.1027**	**0.1110**	100.0	0.0781	0.9999‡	0.4361	0.3772
UB	0.0498‡	0.0618‡	17.8	0.2431	0.9999†	0.5835	0.5594
TD	0.0420	0.0520	17.8	0.6108‡	**0.9999**	0.7838‡	0.7710‡
HKV	0.0498†	0.0611†	17.8	0.3068	0.9998	0.6122	0.5885

- Temporal recommenders less competitive in this dataset (no completely realistic timestamps)
- LIN not very useful
- AIN and MIN are the best metrics to analyze the behavior in terms of temporal novelty

Correctness results

- Can we find a coverage-relevance tradeoff?
- How do correctness metrics compare against other aggregation metrics (F, G)?

Correctness results: MovieLens

σ_τ	P	USC	ISC	F_1	F_2	$F_{0.5}$	$G_{1,1}$	$G_{1,2}$	$G_{2,1}$	UC	RUC	IC	RIC
–	0.093	**100.0**	22.7	0.170	0.338	0.113	0.304	0.453	0.205	0.093	0.093	0.001	0.009
0.82	**0.326**	28.2	9.1	0.303	0.290	**0.316**	0.303	0.296	0.311	0.100	0.094	0.001	0.006
0.84	0.283	59.0	15.1	**0.382**	0.484	0.316	0.408	0.462	**0.361**	0.174	0.170	0.002	0.011
0.86	0.214	80.9	19.6	0.338	**0.520**	0.251	**0.416**	0.519	0.333	**0.177**	**0.176**	**0.002**	0.012
0.88	0.181	95.6	22.2	0.304	0.514	0.216	0.415	**0.548**	0.315	0.176	0.176	0.002	**0.013**
0.90	0.165	99.5	24.8	0.283	0.495	0.198	0.405	0.546	0.300	0.165	0.165	0.002	0.013
0.92	0.156	100.0	26.0	0.269	0.480	0.187	0.395	0.538	0.289	0.156	0.156	0.002	0.012
0.94	0.145	100.0	27.3	0.254	0.459	0.175	0.381	0.526	0.276	0.145	0.145	0.002	0.011
0.96	0.139	100.0	28.2	0.245	0.447	0.168	0.373	0.518	0.269	0.139	0.139	0.002	0.011
0.98	0.133	100.0	**28.6**	0.235	0.435	0.161	0.365	0.511	0.261	0.133	0.133	0.002	0.011

Correctness results: MovieLens

σ_τ	P	USC	ISC	F_1	F_2	$F_{0.5}$	$G_{1,1}$	$G_{1,2}$	$G_{2,1}$	UC	RUC	IC	RIC
–	0.093	**100.0**	22.7	0.170	0.338	0.113	0.304	0.453	0.205	0.093	0.093	0.001	0.009
0.82	**0.326**	28.2	9.1	0.303	0.290	**0.316**	0.303	0.296	0.311	0.100	0.094	0.001	0.006
0.84	0.283	59.0	15.1	**0.382**	0.484	0.316	0.408	0.462	**0.361**	0.174	0.170	0.002	0.011
0.86	0.214	80.9	19.6	0.338	**0.520**	0.251	**0.416**	0.519	0.333	**0.177**	**0.176**	**0.002**	0.012
0.88	0.181	95.6	22.2	0.304	0.514	0.216	0.415	**0.548**	0.315	0.176	0.176	0.002	**0.013**
0.90	0.165	99.5	24.8	0.283	0.495	0.198	0.405	0.546	0.300	0.165	0.165	0.002	0.013
0.92	0.156	100.0	26.0	0.269	0.480	0.187	0.395	0.538	0.289	0.156	0.156	0.002	0.012
0.94	0.145	100.0	27.3	0.254	0.459	0.175	0.381	0.526	0.276	0.145	0.145	0.002	0.011
0.96	0.139	100.0	28.2	0.245	0.447	0.168	0.373	0.518	0.269	0.139	0.139	0.002	0.011
0.98	0.133	100.0	**28.6**	0.235	0.435	0.161	0.365	0.511	0.261	0.133	0.133	0.002	0.011

- Not obvious tradeoff between coverage (USC) and precision (P)

Correctness results: MovieLens

σ_τ	P	USC	ISC	F_1	F_2	$F_{0.5}$	$G_{1,1}$	$G_{1,2}$	$G_{2,1}$	UC	RUC	IC	RIC
–	0.093	**100.0**	22.7	0.170	0.338	0.113	0.304	0.453	0.205	0.093	0.093	0.001	0.009
0.82	**0.326**	28.2	9.1	0.303	0.290	**0.316**	0.303	0.296	0.311	0.100	0.094	0.001	0.006
0.84	0.283	59.0	15.1	**0.382**	0.484	0.316	0.408	0.462	**0.361**	0.174	0.170	0.002	0.011
0.86	0.214	80.9	19.6	0.338	**0.520**	0.251	**0.416**	0.519	0.333	**0.177**	**0.176**	**0.002**	0.012
0.88	0.181	95.6	22.2	0.304	0.514	0.216	0.415	**0.548**	0.315	0.176	0.176	0.002	**0.013**
0.90	0.165	99.5	24.8	0.283	0.495	0.198	0.405	0.546	0.300	0.165	0.165	0.002	0.013
0.92	0.156	100.0	26.0	0.269	0.480	0.187	0.395	0.538	0.289	0.156	0.156	0.002	0.012
0.94	0.145	100.0	27.3	0.254	0.459	0.175	0.381	0.526	0.276	0.145	0.145	0.002	0.011
0.96	0.139	100.0	28.2	0.245	0.447	0.168	0.373	0.518	0.269	0.139	0.139	0.002	0.011
0.98	0.133	100.0	**28.6**	0.235	0.435	0.161	0.365	0.511	0.261	0.133	0.133	0.002	0.011

- Not obvious tradeoff between coverage (USC) and precision (P)
- F_1 and $G_{2,1}$ are too sensitive to the precision value ($\sigma_\tau = 0.84$)

Correctness results: MovieLens

σ_τ	P	USC	ISC	F_1	F_2	$F_{0.5}$	$G_{1,1}$	$G_{1,2}$	$G_{2,1}$	UC	RUC	IC	RIC
–	0.093	**100.0**	22.7	0.170	0.338	0.113	0.304	0.453	0.205	0.093	0.093	0.001	0.009
0.82	**0.326**	28.2	9.1	0.303	0.290	**0.316**	0.303	0.296	0.311	0.100	0.094	0.001	0.006
0.84	0.283	59.0	15.1	**0.382**	0.484	0.316	0.408	0.462	**0.361**	0.174	0.170	0.002	0.011
0.86	0.214	80.9	19.6	0.338	**0.520**	0.251	**0.416**	0.519	0.333	**0.177**	**0.176**	**0.002**	0.012
0.88	0.181	95.6	22.2	0.304	0.514	0.216	0.415	**0.548**	0.315	0.176	0.176	0.002	**0.013**
0.90	0.165	99.5	24.8	0.283	0.495	0.198	0.405	0.546	0.300	0.165	0.165	0.002	0.013
0.92	0.156	100.0	26.0	0.269	0.480	0.187	0.395	0.538	0.289	0.156	0.156	0.002	0.012
0.94	0.145	100.0	27.3	0.254	0.459	0.175	0.381	0.526	0.276	0.145	0.145	0.002	0.011
0.96	0.139	100.0	28.2	0.245	0.447	0.168	0.373	0.518	0.269	0.139	0.139	0.002	0.011
0.98	0.133	100.0	**28.6**	0.235	0.435	0.161	0.365	0.511	0.261	0.133	0.133	0.002	0.011

- Not obvious tradeoff between coverage (USC) and precision (P)
- F_1 and $G_{2,1}$ are too sensitive to the precision value ($\sigma_\tau = 0.84$)
- Best one according to UC: $\sigma_\tau = 0.86$

Correctness results: MovieLens

σ_τ	P	USC	ISC	F_1	F_2	$F_{0.5}$	$G_{1,1}$	$G_{1,2}$	$G_{2,1}$	UC	RUC	IC	RIC
–	0.093	**100.0**	22.7	0.170	0.338	0.113	0.304	0.453	0.205	0.093	0.093	0.001	0.009
0.82	**0.326**	28.2	9.1	0.303	0.290	**0.316**	0.303	0.296	0.311	0.100	0.094	0.001	0.006
0.84	0.283	59.0	15.1	**0.382**	0.484	0.316	0.408	0.462	**0.361**	0.174	0.170	0.002	0.011
0.86	0.214	80.9	19.6	0.338	**0.520**	0.251	**0.416**	0.519	0.333	**0.177**	**0.176**	**0.002**	0.012
0.88	0.181	95.6	22.2	0.304	0.514	0.216	0.415	**0.548**	0.315	0.176	0.176	0.002	**0.013**
0.90	0.165	99.5	24.8	0.283	0.495	0.198	0.405	0.546	0.300	0.165	0.165	0.002	0.013
0.92	0.156	100.0	26.0	0.269	0.480	0.187	0.395	0.538	0.289	0.156	0.156	0.002	0.012
0.94	0.145	100.0	27.3	0.254	0.459	0.175	0.381	0.526	0.276	0.145	0.145	0.002	0.011
0.96	0.139	100.0	28.2	0.245	0.447	0.168	0.373	0.518	0.269	0.139	0.139	0.002	0.011
0.98	0.133	100.0	**28.6**	0.235	0.435	0.161	0.365	0.511	0.261	0.133	0.133	0.002	0.011

- Not obvious tradeoff between coverage (USC) and precision (P)
- F_1 and $G_{2,1}$ are too sensitive to the precision value ($\sigma_\tau = 0.84$)
- Best one according to UC: $\sigma_\tau = 0.86$
- However, these values decrease recommendation novelty and diversity

Outline

1. Recommender Systems
2. Freshness
3. Correctness
4. Experiments
5. Conclusions and future work

Conclusions

- Freshness
  - We introduced the temporal dimensions in the definition of a family of novelty models
  - The proposed metric works as expected although it can be affected by biases in the data
  - For more information, see **Sánchez and Bellogín (2018)**.

Conclusions

- Freshness
  - We introduced the temporal dimensions in the definition of a family of novelty models
  - The proposed metric works as expected although it can be affected by biases in the data
  - For more information, see **Sánchez and Bellogín (2018)**.
- Correctness
  - We have proposed a set of metrics on the assumption that it is better to avoid a recommendation rather than providing a bad recommendation
  - We have shown that it is not easy to balance precision, coverage, and novelty and diversity
  - For more information, see **Mesas and Bellogín (2017)**

Future work

- Freshness
  - Freshness analysis could favor new possibilities to produce time-aware recommendation whenever relevance is not the only important dimension
  - These temporal models could also be applied in online recommender systems, such as news recommendation.

Future work

- Freshness
  - Freshness analysis could favor new possibilities to produce time-aware recommendation whenever relevance is not the only important dimension
  - These temporal models could also be applied in online recommender systems, such as news recommendation.
- Correctness
  - Extend correctness to combine other evaluation dimensions (freshness, novelty, and diversity)
  - Analyze the bad recommendations that we may provide to the user from a more formal point of view

New approaches for evaluation: correctness and freshness

Pablo Sánchez Rus M. Mesas Alejandro Bellogín

Universidad Autónoma de Madrid
Escuela Politécnica Superior
Departamento de Ingeniería Informática

V Congreso Español de
Recuperación de Información (CERI 2018)

Thank you

Freshness: Datasets

Dataset	Users	Items	Ratings	Density	Scale	Date range
Ep (2-core)	22,556	15,196	75,533	0.022%	[1, 5]	Jan 2001 - Nov 2013
ML	138,493	26,744	20,000,263	0.540%	[0.5, 5]	Jan 1995 - Mar 2015
MT (5-core)	15,411	8,443	518,558	0.398%	[0, 10]	Feb 2013 - Apr 2017

- MovieTweetings and Movielens20M are from the movie domain
- Epinions dataset contains purchases of different products

Freshness: Datasets

Dataset	Users	Items	Ratings	Density	Scale	Date range
Ep (2-core)	22,556	15,196	75,533	0.022%	[1,5]	Jan 2001 - Nov 2013
ML	138,493	26,744	20,000,263	0.540%	[0.5,5]	Jan 1995 - Mar 2015
MT (5-core)	15,411	8,443	518,558	0.398%	[0,10]	Feb 2013 - Apr 2017

- MovieTweetings and Movielens20M are from the movie domain
- Epinions dataset contains purchases of different products

Freshness: Datasets

Dataset	Users	Items	Ratings	Density	Scale	Date range
Ep (2-core)	22,556	15,196	75,533	0.022%	[1, 5]	Jan 2001 - Nov 2013
ML	138,493	26,744	20,000,263	0.540%	[0.5, 5]	Jan 1995 - Mar 2015
MT (5-core)	15,411	8,443	518,558	0.398%	[0, 10]	Feb 2013 - Apr 2017

- MovieTweetings and Movielens20M are from the movie domain
- Epinions dataset contains purchases of different products
- All datasets contain timestamps

Freshness: Datasets

Dataset	Users	Items	Ratings	Density	Scale	Date range
Ep (2-core)	22,556	15,196	75,533	0.022%	[1, 5]	Jan 2001 - Nov 2013
ML	138,493	26,744	20,000,263	0.540%	[0.5, 5]	Jan 1995 - Mar 2015
MT (5-core)	15,411	8,443	518,558	0.398%	[0, 10]	Feb 2013 - Apr 2017

- MovieTweetings and Movielens20M are from the movie domain
- Epinions dataset contains purchases of different products
- All datasets contain timestamps
- All metrics @5

Freshness: Datasets

Dataset	Users	Items	Ratings	Density	Scale	Date range
Ep (2-core)	22,556	15,196	75,533	0.022%	[1, 5]	Jan 2001 - Nov 2013
ML	138,493	26,744	20,000,263	0.540%	[0.5, 5]	Jan 1995 - Mar 2015
MT (5-core)	15,411	8,443	518,558	0.398%	[0, 10]	Feb 2013 - Apr 2017

- MovieTweetings and Movielens20M are from the movie domain
- Epinions dataset contains purchases of different products
- All datasets contain timestamps
- All metrics @5
- Relevance thresholds of 5 for Ep and ML and 9 for MT

Freshness: Recommenders

- Non-personalized: Rnd, Pop, IdAsc, IdDec
- Personalized: UB, HKV (MF)
- Personalized and time/sequence aware: TD (UB)
- Skylines (perfect recommenders):
    - SkyPerf: returns the test set
    - SkyFresh: optimizes one of the freshness models (LIN)

Freshness: Recommenders

- Non-personalized: Rnd, Pop, IdAsc, IdDec
- Personalized: UB, HKV (MF)[1]
- Personalized and time/sequence aware: TD (UB)
- Skylines (perfect recommenders):
  - SkyPerf: returns the test set
  - SkyFresh: optimizes one of the freshness models (LIN)

[1] Hu et al. (2008)

Freshness: Recommenders

- Non-personalized: Rnd, Pop, IdAsc, IdDec
- Personalized: UB, HKV (MF)
- Personalized and time/sequence aware: TD (UB)[1]
- Skylines (perfect recommenders):
  - SkyPerf: returns the test set
  - SkyFresh: optimizes one of the freshness models (LIN)

[1] Based on Ding and Li (2005)

Results: MovieTweetings

Algorithm	P	NDCG	USC	No relevance FIN	LIN	AIN	MIN
Rnd	0.0002	0.0003	**100.0**	0.1693	0.8473	0.4435	0.4086
IdAsc	0.0004	0.0003	100.0‡	0.1729	0.8873	0.5485	0.5938†
IdDec	0.0005	0.0004	100.0†	**0.9628**	0.9800	**0.9688**	**0.9669**
Pop	0.0028	0.0023	100.0	0.1499	0.9921	0.2534	0.2074
UB	0.0104†	0.0120†	78.5	0.4902†	0.9951‡	0.5937†	0.5657
TD	**0.0264**	**0.0337**	78.5	0.8487‡	**0.9988**	0.9298‡	0.9282‡
HKV	0.0150‡	0.0190‡	78.5	0.4131	0.9939†	0.5935	0.5621

- Higher coverage in personalized recommenders than before (shorter time-range)
- Item ordering bias (items with higher id are more fresh)
- Temporal recommender competitive when using more realistic timestamps

Correctness: Datasets

Dataset	Users	Items	Ratings	Density	Scale
Movielens100K	943	1681	$100,000$	6.3%	$[1,5]$
Jester	$59,132$	150	$1,710,677$	19.28%	$[0,20]$
Movielens1M	$6,040$	$3,883$	$1,000,209$	4.26%	$[1,5]$

- Movielens100K and Movielens1M are from the movie domain
- Jester is a jokes dataset
- All metrics @5

References I

Ding, Y. and Li, X. (2005). Time weight collaborative filtering. In *CIKM*, pages 485–492. ACM.

Gunawardana, A. and Shani, G. (2015). Evaluating recommender systems. In *Recommender Systems Handbook*, pages 265–308. Springer.

Herlocker, J. L., Konstan, J. A., Terveen, L. G., and Riedl, J. (2004). Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.*, 22(1):5–53.

Hu, Y., Koren, Y., and Volinsky, C. (2008). Collaborative filtering for implicit feedback datasets. In *ICDM*, pages 263–272. IEEE Computer Society.

References II

Mesas, R. M. and Bellogín, A. (2017). Evaluating decision-aware recommender systems. In Cremonesi, P., Ricci, F., Berkovsky, S., and Tuzhilin, A., editors, *Proceedings of the Eleventh ACM Conference on Recommender Systems, RecSys 2017, Como, Italy, August 27-31, 2017*, pages 74–78. ACM.

Peñas, A. and Rodrigo, Á. (2011). A simple measure to assess non-response. In Lin, D., Matsumoto, Y., and Mihalcea, R., editors, *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA*, pages 1415–1424. The Association for Computer Linguistics.

References III

Sánchez, P. and Bellogín, A. (2018). Time-aware novelty metrics for recommender systems. In Pasi, G., Piwowarski, B., Azzopardi, L., and Hanbury, A., editors, *Advances in Information Retrieval - 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings*, volume 10772 of *Lecture Notes in Computer Science*, pages 357–370. Springer.

Vargas, S. and Castells, P. (2011). Rank and relevance in novelty and diversity metrics for recommender systems. In *RecSys*, pages 109–116. ACM.