Time-Aware Novelty Metrics for Recommender Systems

Pablo Sánchez Alejandro Bellogín

Universidad Autónoma de Madrid
Escuela Politécnica Superior
Departamento de Ingeniería Informática

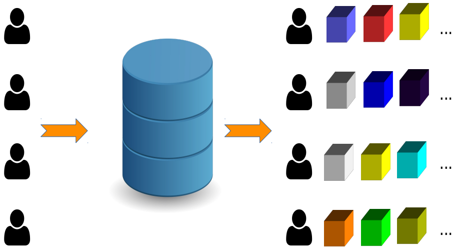
European Conference on Information Retrieval, 2018

Outline

Outline

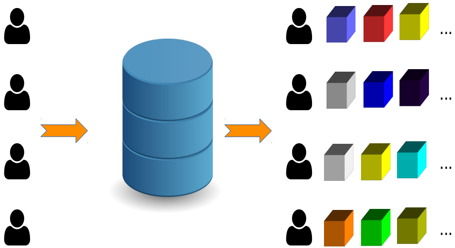
1. Recommender Systems
2. Time-Aware Novelty Metrics for Recommender Systems
3. Experiments
4. Conclusions and future work

Recommender Systems



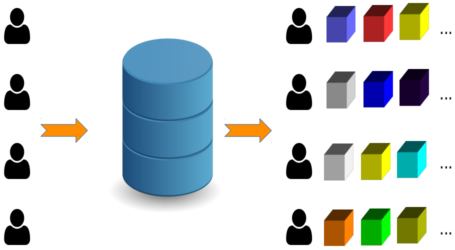
- Suggest **new items** to users based on their tastes and needs

Recommender Systems



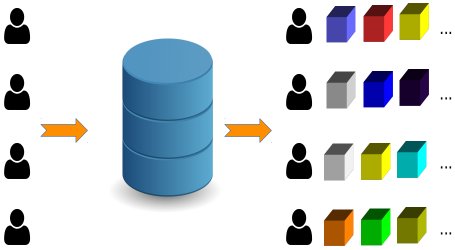
- Suggest **new items** to users based on their tastes and needs
- Measure the **quality** of recommendations. How?

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

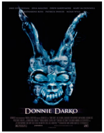
Recommender Systems



- 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 the **temporal dimension**

Different notions of quality

R_1

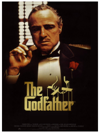


(2001)

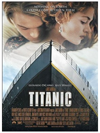
(1994)

(1994)

R_2



(1972)



(1997)

(1993)

R_3

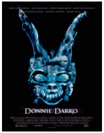
(2018)

(2017)

(2016)

Different notions of quality

R_1

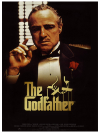


(2001)

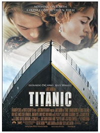
(1994)

(1994)

R_2



(1972)



(1997)

(1993)

R_3

(2018)

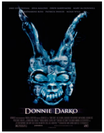
(2017)

(2016)

- Best in Relevance?

Different notions of quality

R_1

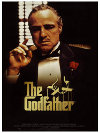


(2001)

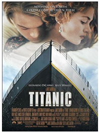
(1994)

(1994)

R_2



(1972)



(1997)

(1993)

R_3

(2018)

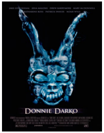
(2017)

(2016)

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

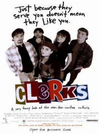
Different notions of quality

R_1



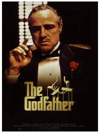
(2001)

(1994)

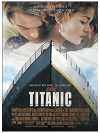


(1994)

R_2



(1972)



(1997)

(1993)

R_3

(2018)

(2017)

(2016)

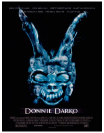
- Best in Relevance?
  - $R_2 > R_1 > R_3$
- Best in Novelty?

Different notions of quality

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

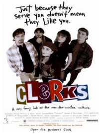
Different notions of quality

R_1



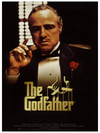
(2001)

(1994)

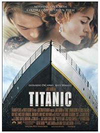


(1994)

R_2



(1972)



(1997)

(1993)

R_3

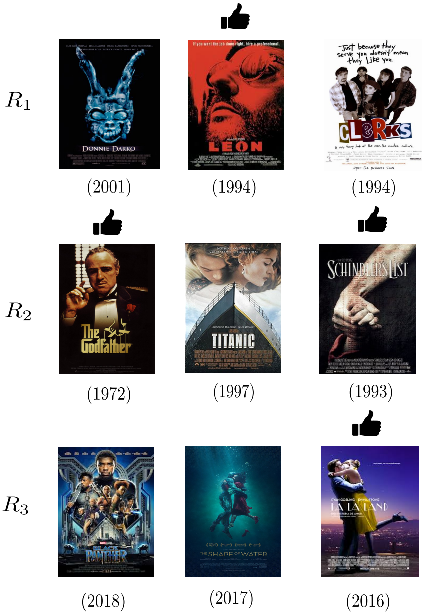
(2018)

(2017)

(2016)

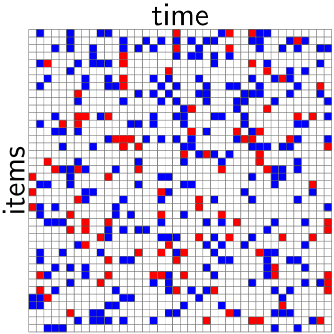
- Best in Relevance?
  - $R_2 > R_1 > R_3$
- Best in Novelty?
  - $R_1 > R_3 > R_2$
- Best in **Freshness**?

Different notions of quality

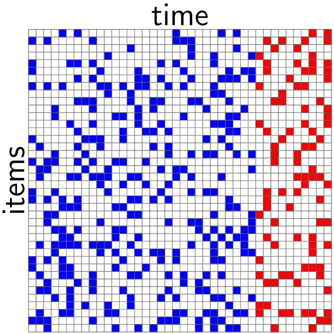


- Best in Relevance?
  - $R_2 > R_1 > R_3$
- Best in Novelty?
  - $R_1 > R_3 > R_2$
- Best in **Freshness**?
  - $R_3 > R_1 > R_2$

Types of data splitting

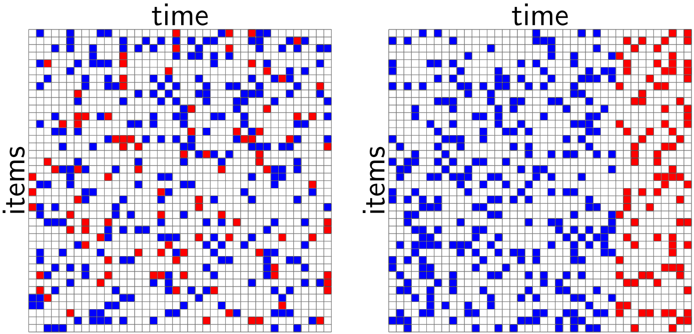


Random split



Temporal split

Types of data splitting

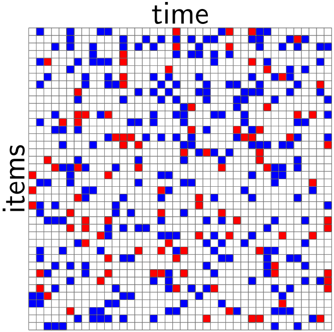


Random split

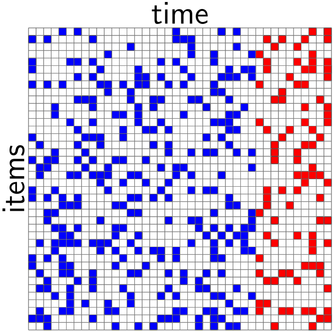
Temporal split

- Random splitting has been the most extended way to test recommender systems

Types of data splitting



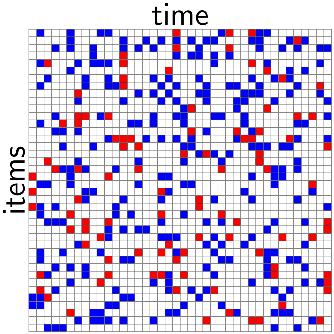
Random split



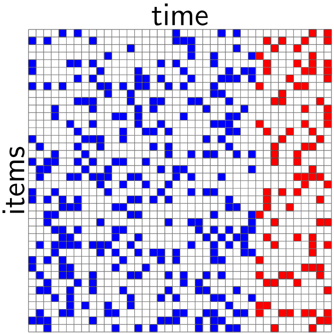
Temporal split

- Random splitting has been the most extended way to test recommender systems
- Temporal splitting is becoming more important

Types of data splitting



Random split



Temporal split

- Random splitting has been the most extended way to test recommender systems
- Temporal splitting is becoming more important
  - Hence, time should also be incorporated in evaluation metrics

Outline

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)

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

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

Preliminaries

- Framework proposed in Vargas and Castells (2011)

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

- When using $\mathrm{nov}(i_n \mid \theta) = (1 - p(\mathrm{seen}|i))$ we obtain the expected popularity complement (EPC) metric

Preliminaries

- Framework proposed in Vargas and Castells (2011)

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

- When using $\text{nov}(i_n \mid \theta) = (1 - p(\text{seen}|i))$ we obtain the expected popularity complement (EPC) metric

- However, all the metrics derived from this framework are *time-agnostic*

Preliminaries

- Framework proposed in Vargas and Castells (2011)

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

- When using $\mathrm{nov}(i_n \mid \theta) = (1 - p(\mathrm{seen}|i))$ we obtain the expected popularity complement (EPC) metric
- However, all the metrics derived from this framework 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:

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

- 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.

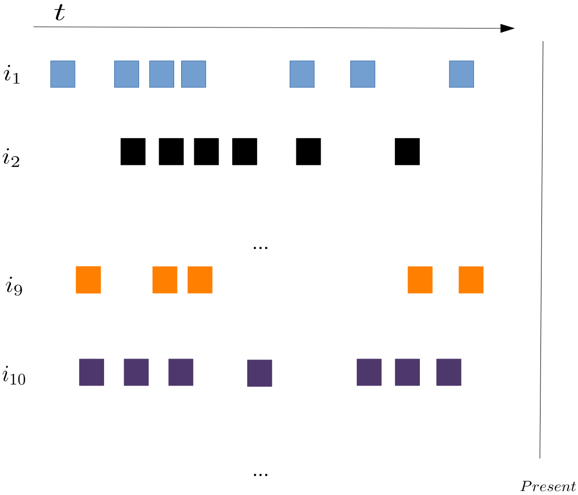
Time-Aware Novelty Metrics

- 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)

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)

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)

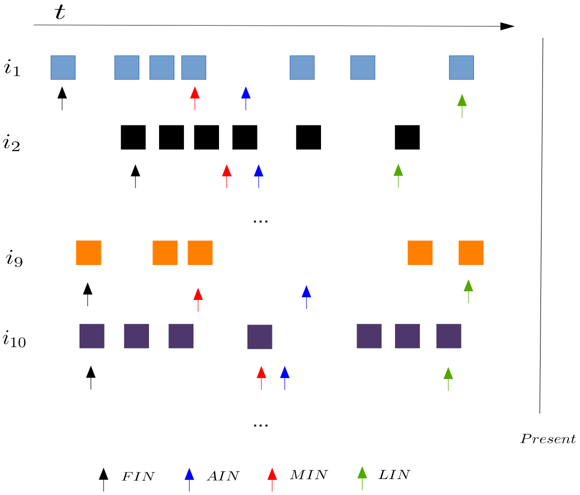
Modeling time profiles for items

- 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

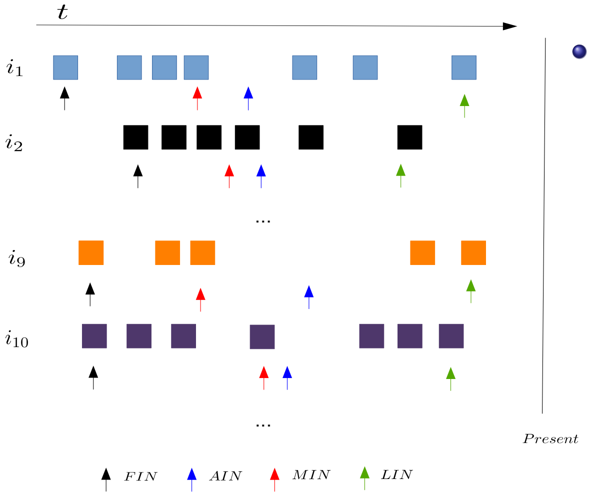
- How can we aggregate the temporal representation?
- We explored four possibilities:
  - Take the first interaction (FIN)
  - Take the last interaction (LIN)
  - Take the average of the ratings times (AIN)
  - Take the median of the ratings times (MIN)
- Each case defines a function $f(\theta_t(i))$

Modeling time profiles for items: an example



Modeling time profiles for items: an example

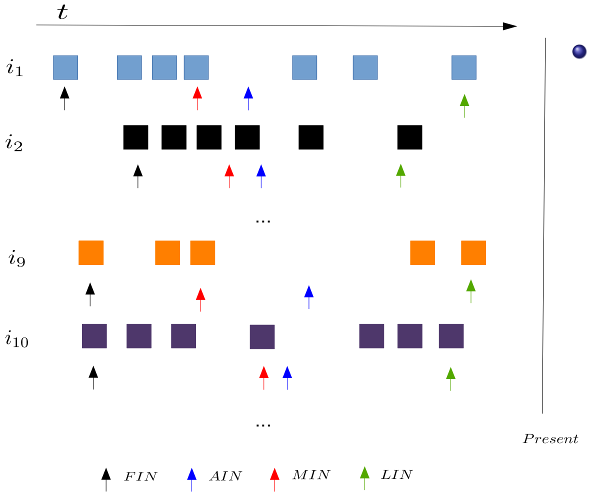
- Which model represents better the freshness of the items?



- FIN?

Modeling time profiles for items: an example

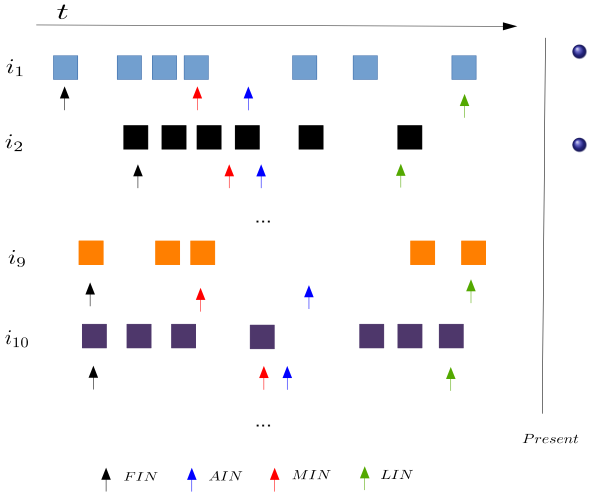
- Which model represents better the freshness of the items?



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

Modeling time profiles for items: an example

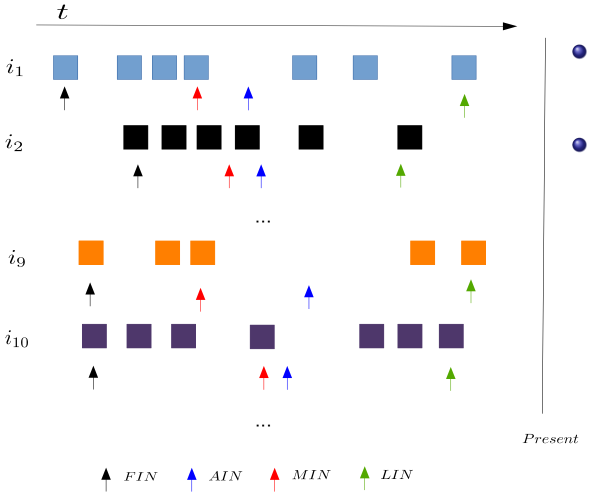
- Which model represents better the freshness of the items?



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

Modeling time profiles for items: an example

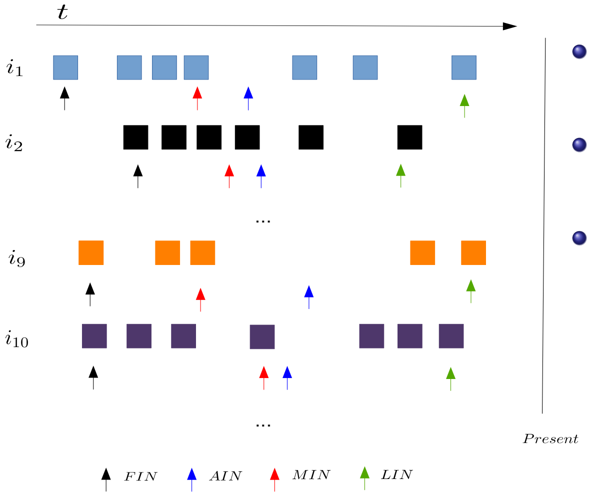
- Which model represents better the freshness of the items?



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

Modeling time profiles for items: an example

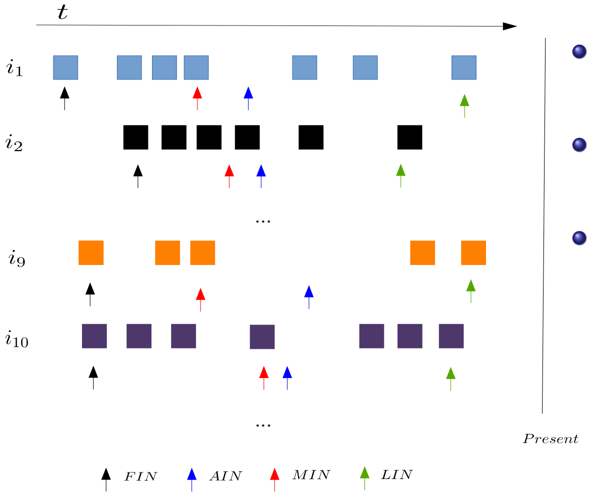
- Which model represents better the freshness of the items?



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

Modeling time profiles for items: an example

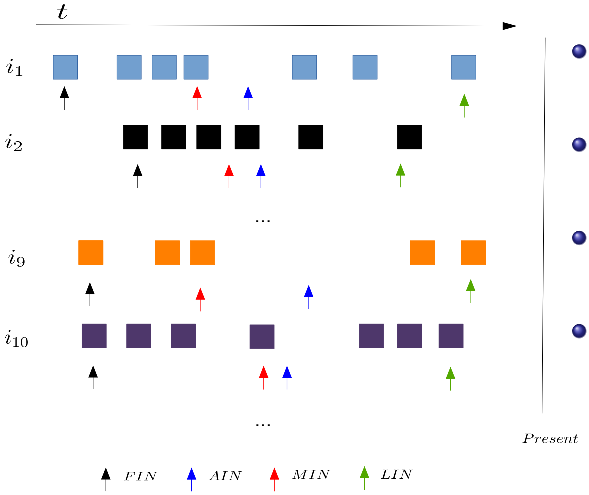
- Which model represents better the freshness of the items?



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

Modeling time profiles for items: an example

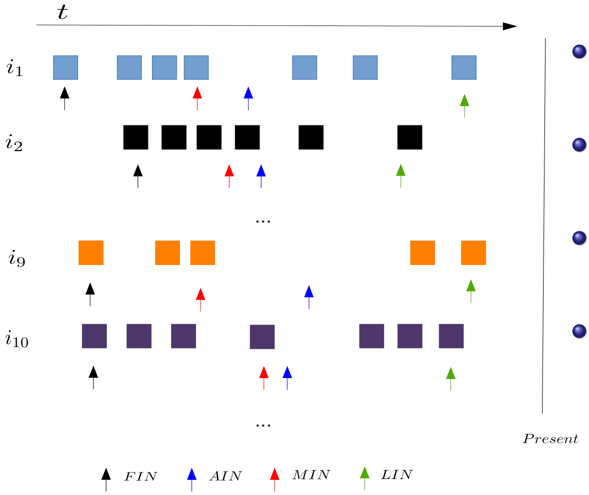
- Which model represents better the freshness of the items?



- 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?

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?
  - $i_9 > i_{10} > i_2 > i_1$

Integration in the framework

- The proposed models are not suitable for the probabilistic framework:

$$m(R_u \mid \theta_t) = C \sum_{i_n \in R_u} \text{disc}(n) p(rel \mid i_n, u) \boxed{\text{nov}(i_n \mid \theta_t)} \tag{3}$$

Integration in the framework

- The proposed models are not suitable for the probabilistic framework:

$$m(R_u \mid \theta_t) = C \sum_{i_n \in R_u} \mathrm{disc}(n) p(rel \mid i_n, u) \boxed{\mathrm{nov}(i_n \mid \theta_t)} \tag{3}$$

- We apply a normalization step: either min-max normalization or dividing by the largest timestamp

Integration in the framework

- The proposed models are not suitable for the probabilistic framework:

$$m(R_u \mid \theta_t) = C \sum_{i_n \in R_u} \mathrm{disc}(n) p(rel \mid i_n, u) \boxed{\mathrm{nov}(i_n \mid \theta_t)} \quad (3)$$

- We apply a normalization step: either min-max normalization or dividing by the largest timestamp

$$\mathrm{nov}^{f,n}(i \mid \theta_t) = n(f(\theta_t(i)), \theta_t) \quad (4)$$

Experiments

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

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

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

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

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

Datasets: rating temporal activity

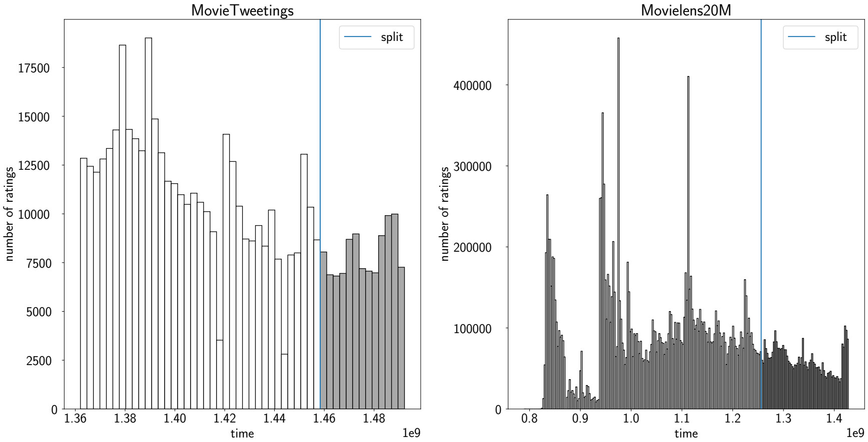


Figure: Rating histogram evolution in MovieTweetings (left) and Movielens20M (right). Temporal split with 80% of older ratings to train the recommenders

Recommenders

- Non-personalized: Rnd, Pop, IdAsc, IdDec

Recommenders

- Non-personalized: Rnd, Pop, IdAsc, IdDec
- Personalized: UB, HKV (MF)[1]

[1] Hu et al. (2008)

Recommenders

- Non-personalized: Rnd, Pop, IdAsc, IdDec
- Personalized: UB, HKV (MF)
- Personalized and time/sequence aware: TD (UB)[1]

[1] Based on Ding and Li (2005)

Recommenders

- Non-personalized: Rnd, Pop, IdAsc, IdDec
- Personalized: UB, HKV (MF)
- Personalized and time/sequence aware: TD (UB)
- Skylines (perfect recommenders):

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

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)

Results: MovieLens

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
SkyPerf	**0.7094**	**0.8396**	99.7	0.6069†	0.9993	0.7764†	0.7618†
SkyFresh	0.0027	0.0027	100.0	0.4999	**1.0000**	0.7236	0.7026

Results: MovieLens

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
SkyPerf	**0.7094**	**0.8396**	99.7	0.6069†	0.9993	0.7764†	0.7618†
SkyFresh	0.0027	0.0027	100.0	0.4999	**1.0000**	0.7236	0.7026

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

Results: MovieLens

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
SkyPerf	**0.7094**	**0.8396**	99.7	0.6069†	0.9993	0.7764†	0.7618†
SkyFresh	0.0027	0.0027	100.0	0.4999	**1.0000**	0.7236	0.7026

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

Results: MovieLens

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
SkyPerf	**0.7094**	**0.8396**	99.7	0.6069†	0.9993	0.7764†	0.7618†
SkyFresh	0.0027	0.0027	100.0	0.4999	**1.0000**	0.7236	0.7026

- Relevance metrics (Precision 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)

Results: MovieLens

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
SkyPerf	**0.7094**	**0.8396**	99.7	0.6069†	0.9993	0.7764†	0.7618†
SkyFresh	0.0027	0.0027	100.0	0.4999	**1.0000**	0.7236	0.7026

- Relevance metrics (Precision 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

Results: Popularity bias

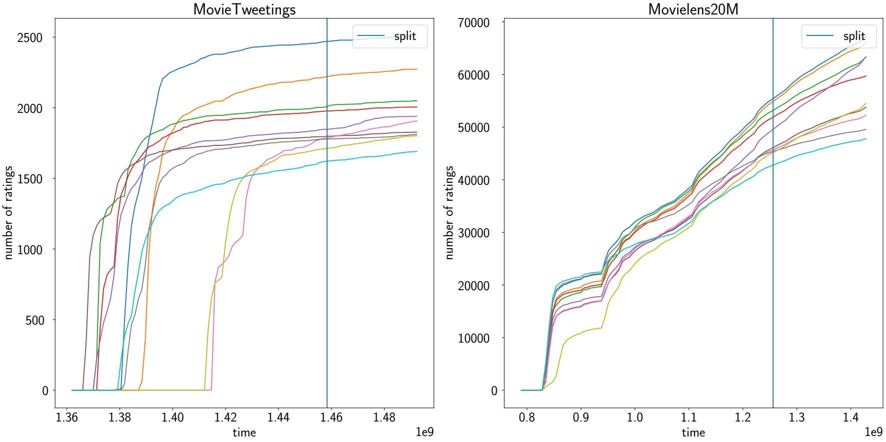


Figure: Top 10 most popular items in the training set of each dataset: MovieTweetings (left) and MovieLens (right).

Results: MovieLens

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
SkyPerf	**0.7094**	**0.8396**	99.7	0.6069†	0.9993	0.7764†	0.7618†
SkyFresh	0.0027	0.0027	100.0	0.4999	**1.0000**	0.7236	0.7026

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

Results: MovieLens

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
SkyPerf	**0.7094**	**0.8396**	99.7	0.6069†	0.9993	0.7764†	0.7618†
SkyFresh	0.0027	0.0027	100.0	0.4999	**1.0000**	0.7236	0.7026

- Temporal recommenders less competitive in this dataset (no completely realistic timestamps)
- Skyline does not achieve maximum performance results (due to evaluation methodology)

Results: MovieLens

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
SkyPerf	**0.7094**	**0.8396**	99.7	0.6069†	0.9993	0.7764†	0.7618†
SkyFresh	0.0027	0.0027	100.0	0.4999	**1.0000**	0.7236	0.7026

- Temporal recommenders less competitive in this dataset (no completely realistic timestamps)
- Skyline does not achieve maximum performance results (due to evaluation methodology)
- LIN not very useful

Results: MovieLens

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
SkyPerf	**0.7094**	**0.8396**	99.7	0.6069†	0.9993	0.7764†	0.7618†
SkyFresh	0.0027	0.0027	100.0	0.4999	**1.0000**	0.7236	0.7026

- Temporal recommenders less competitive in this dataset (no completely realistic timestamps)
- Skyline does not achieve maximum performance results (due to evaluation methodology)
- LIN not very useful
- AIN and MIN are the best metrics to analyze the behavior in terms of temporal novelty

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
SkyPerf	**0.3468**	**0.5374**	81.6	0.4262	0.9686	0.6514	0.6289
SkyFresh	0.0037	0.0041	100.0	0.6715†	**1.0000**	0.8072†	0.7924†

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
SkyPerf	**0.3468**	**0.5374**	81.6	0.4262	0.9686	0.6514	0.6289
SkyFresh	0.0037	0.0041	100.0	0.6715†	**1.0000**	0.8072†	0.7924†

- Higher coverage in personalized recommenders than before (shorter time-range)

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
SkyPerf	**0.3468**	**0.5374**	81.6	0.4262	0.9686	0.6514	0.6289
SkyFresh	0.0037	0.0041	100.0	0.6715†	**1.0000**	0.8072†	0.7924†

- Higher coverage in personalized recommenders than before (shorter time-range)
- Item ordering bias (items with higher id are more fresh)

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
SkyPerf	**0.3468**	**0.5374**	81.6	0.4262	0.9686	0.6514	0.6289
SkyFresh	0.0037	0.0041	100.0	0.6715†	**1.0000**	0.8072†	0.7924†

- 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

Outline

Conclusions and future work

- We introduced the temporal dimensions in the definition of a family of novelty models

Conclusions and future work

- 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

Conclusions and future work

- 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
- This approach could favor new possibilities to produce time-aware recommendation whenever relevance is not the only important dimension

Conclusions and future work

- 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
- This approach 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

Conclusions and future work

- 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
- This approach 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
- Source code and more details to reproduce the experiments in `https://bitbucket.org/PabloSanchezP/timeawarenoveltymetrics`

Time-Aware Novelty Metrics for Recommender Systems

Pablo Sánchez Alejandro Bellogín

Universidad Autónoma de Madrid
Escuela Politécnica Superior
Departamento de Ingeniería Informática

European Conference on Information Retrieval, 2018

Thank you

https://bitbucket.org/PabloSanchezP/timeawarenoveltymetrics

Other approximations related to our freshness metric

- Forgotten Curve in Hu and Ogihara (2011)
  - Exponential function taking into account the number of times the song was played and the distance from the present time to the last time the song was played

Other approximations related to our freshness metric

- Forgotten Curve in Hu and Ogihara (2011)
  - Exponential function taking into account the number of times the song was played and the distance from the present time to the last time the song was played
- Overlap between previous recommendation lists in Lathia et al. (2010):
  - Difference between the items that we are recommending and the ones we have previously recommended to the user

Other approximations related to our freshness metric

- Forgotten Curve in Hu and Ogihara (2011)
  - Exponential function taking into account the number of times the song was played and the distance from the present time to the last time the song was played
- Overlap between previous recommendation lists in Lathia et al. (2010):
  - Difference between the items that we are recommending and the ones we have previously recommended to the user
- Similar approach with metadata: Chou et al. (2015)
  - Taking the average of the release dates of the songs

UB vs TD

- The score of every item for a UB is:

$$\hat{s}_{ui} = \sum_{v \in N_u} sim(u, v) \cdot r_{vi} \tag{5}$$

- The score of every item of the TD is:

$$\hat{s}_{ui} = \sum_{v \in N_u} sim(u, v) \cdot r_{vi} \cdot e^{-\lambda(days(t, t(v,i)))} \tag{6}$$

HKV and BPR

- HKV

$$\min_{x*,y*} \sum_{u,i} c_{ui}(p_{ui} - x_u^T y_i)^2 + \lambda(\sum_u ||x_u||^2 + \sum_i ||y_i||^2) \qquad (7)$$

  - where x_u and y_i are the item factors.

- BPRMF
  - It works with triplets $D_s : U \times I \times I$
  - Optimization of $\sum_{(u,i,j)} \log(\sigma(S(i;u) - S(j;u)))$ (BPR-OPT)
  - in BPR-MF $S(i;u) = \sum_f p_{uf} q_{if}$
  - Θ (model parameters) optimization is done by stochastic gradient descent (choosing the triplets randomly)

Metrics

- MAE and RMSE

$$MAE = \frac{1}{|\mathcal{R}_{test}|} \sum_{r_{ui} \in \mathcal{R}_{test}} |g(u,i) - r_{ui}| \tag{8}$$

$$RMSE = \sqrt{\frac{1}{|\mathcal{R}_{test}|} \sum_{r_{ui} \in \mathcal{R}_{test}} (g(u,i) - r_{ui})^2} \tag{9}$$

- Precision

$$Precision = \frac{\text{Relevant items} \cap \text{Retrieved items}}{\text{Retrieved items}} \tag{10}$$

- NDCG

$$NDCG_p = \frac{DCG_p}{IDCG_p} \tag{11}$$

$$DCG_p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2 i} \tag{12}$$

Epinions results

Algorithm	P	NDCG	USC	No relevance FIN	No relevance LIN	No relevance AIN	No relevance MIN	Relevance FIN	Relevance LIN	Relevance AIN	Relevance MIN
Rnd	0.0000	0.0001	**100.0**	0.3812	0.6391	0.4901	0.4753	0.0000	0.0000	0.0000	0.0000
IdAsc	0.0000	0.0000	100.0‡	0.2357	0.5083	0.3599	0.3401	0.0000	0.0000	0.0000	0.0000
IdDec	0.0000	0.0001	100.0†	0.3851	0.5790	0.4766	0.4728	0.0000	0.0000	0.0000	0.0000
Pop	0.0009‡	0.0012†	100.0	0.0788	0.7936	0.2670	0.2152	0.0003	0.0009‡	0.0006‡	0.0005‡
IB	0.0002	0.0005	49.7	0.4567†	0.6705	0.5505	0.5411	0.0001	0.0001	0.0001	0.0001
UB	0.0004	0.0007	49.7	0.3325	0.7625	0.4871	0.4601	0.0001	0.0004	0.0003	0.0003
TD	0.0004	0.0008	49.7	0.6000‡	0.9150‡	**0.7365**	**0.7238**	0.0003†	0.0004	0.0003	0.0003
HKV	0.0006	0.0018‡	50.6	0.2445	0.8808†	0.4366	0.3977	0.0002	0.0006	0.0004	0.0004
BPR	0.0007†	0.0011	50.6	0.1964	0.7917	0.3705	0.3362	0.0004‡	0.0007†	0.0005†	0.0005†
Fossil	0.0002	0.0004	31.1	0.2821	0.7806	0.4527	0.4200	0.0001	0.0001	0.0001	0.0001
SkyPerf	**0.1337**	**0.4441**	66.5	**0.6170**	0.8695	0.7286‡	0.7197‡	**0.2397**	**0.3416**	**0.2845**	**0.2807**
SkyFresh	0.0000	0.0000	100.0	0.4557	**0.9999**	0.6588†	0.5976†	0.0000	0.0000	0.0000	0.0000

Results with meta-data information

Algorithm	No relevance ML	
	Y-*IN	R-FIN
Rnd	0.7707	0.5573
IdAsc	0.8387†	0.0716
IdDec	0.7581	**0.9995**
Pop	0.8227	0.0781
UB	0.8164	0.2431
TD	**0.8822**	0.6108‡
HKV	0.8102	0.3068
SkyPerf	0.8602‡	0.6069†
SkyFresh	0.6305	0.4999

Algorithm	No relevance MT	
	Y-*IN	R-FIN
Rnd	0.8764	0.1693
IdAsc	0.2264	0.1729
IdDec	**0.9907**	**0.9628**
Pop	0.9693	0.1499
UB	0.9745†	0.4902
TD	0.9817‡	0.8487‡
HKV	0.9494	0.4131
SkyPerf	0.9184	0.4262
SkyFresh	0.9689	0.6715†

Results with meta-data information

Algorithm	No relevance ML Y-*IN	No relevance ML R-FIN
Rnd	0.7707	0.5573
IdAsc	0.8387†	0.0716
IdDec	0.7581	**0.9995**
Pop	0.8227	0.0781
UB	0.8164	0.2431
TD	**0.8822**	0.6108‡
HKV	0.8102	0.3068
SkyPerf	0.8602‡	0.6069†
SkyFresh	0.6305	0.4999

Algorithm	No relevance MT Y-*IN	No relevance MT R-FIN
Rnd	0.8764	0.1693
IdAsc	0.2264	0.1729
IdDec	**0.9907**	**0.9628**
Pop	0.9693	0.1499
UB	0.9745†	0.4902
TD	0.9817‡	0.8487‡
HKV	0.9494	0.4131
SkyPerf	0.9184	0.4262
SkyFresh	0.9689	0.6715†

- TD also retrieving fresh items when using metadata

Results with meta-data information

Algorithm	No relevance ML	
	Y-*IN	R-FIN
Rnd	0.7707	0.5573
IdAsc	0.8387†	0.0716
IdDec	0.7581	**0.9995**
Pop	0.8227	0.0781
UB	0.8164	0.2431
TD	**0.8822**	0.6108‡
HKV	0.8102	0.3068
SkyPerf	0.8602‡	0.6069†
SkyFresh	0.6305	0.4999

Algorithm	No relevance MT	
	Y-*IN	R-FIN
Rnd	0.8764	0.1693
IdAsc	0.2264	0.1729
IdDec	**0.9907**	**0.9628**
Pop	0.9693	0.1499
UB	0.9745†	0.4902
TD	0.9817‡	0.8487‡
HKV	0.9494	0.4131
SkyPerf	0.9184	0.4262
SkyFresh	0.9689	0.6715†

- TD also retrieving fresh items when using metadata
- Different behavior between old items (by release date) and items with a high lifespan in both datasets

References I

Chou, S., Yang, Y., and Lin, Y. (2015). Evaluating music recommendation in a real-world setting: On data splitting and evaluation metrics. In *ICME*, pages 1–6. IEEE Computer Society.

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

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

Hu, Y. and Ogihara, M. (2011). Nextone player: A music recommendation system based on user behavior. In *ISMIR*, pages 103–108. University of Miami.

Lathia, N., Hailes, S., Capra, L., and Amatriain, X. (2010). Temporal diversity in recommender systems. In *SIGIR*, pages 210–217. ACM.

References II

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