IIC3666 - Sistemas Recomendadores

# Widespread Flaws in Offline Evaluation of Recommender Systems

Balázs Hidasi and Ádám Tibor Czapp

Integrantes:

- Nicolas Guzman
- Valentina Nuñez
- Alonso Zamorano
- Diego Jiménez

# Contexto

Contexto

# A/B Testing

- Aproximacion limitada, debido a KPIs
- Division de trafico
- Leak de información
- Sesgos
- Costos y lentitud
- Falta de reproducibilidad

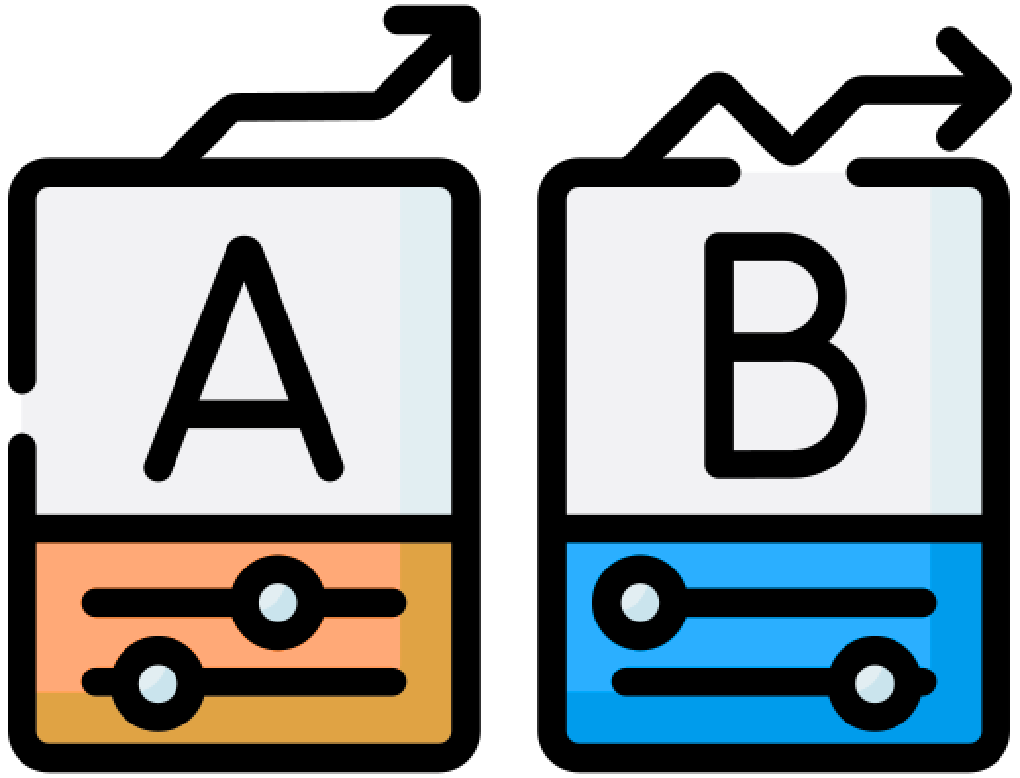



image: Flaticon.com

**A/B Testing**

# Evaluacion Offline

## 4 errores

- Dataset-task Mismatch
- Overzealous preprocessing
- Information leaking through time
- Negative sampling during testing

# Testing

- Next item prediction
- Session-based Recomender
- Behaviour prediction (eventos + interacciones con recomendador)
- recall@N y MRR@N
- GRU4Rec

# Datasets

**No Secuenciales (Rating):**

- Amazon (beauty)
- Movie Lens
- Steam
- Yelp

**Secuenciales:**

- Rees46
- Coveo (artificial)
- Retail Rocket

# Dataset-Task Mismatch

Dataset-Task Mismatch

# Ejemplos de errores

- Rating como feedback implicito
- Recomendación secuencial erronea
- Colisiones de eventos en secuencias

Dataset-Task Mismatch

# Datasets

Table 1. Basic statistics of train/test splits and event collision rate of the datasets

| Dataset | Training set #Events | Training set #Sequences | Training set #Days | Test set #Events | Test set #Sequences | Test set #Days | #Items | Event time collisions Proportion | Event time collisions Event% |
|---|---|---|---|---|---|---|---|---|---|
| Amazon (Beauty) | 724,440 | 215,595 | 4,907 | 30,191 | 11,452 | 56 | 38,606 | 31.89% | 33.03% |
| MovieLens10M | 9,861,612 | 69,141 | 5,054 | 99,022 | 737 | 56 | 10,066 | 17.83% | 27.33% |
| Steam | 4,856,479 | 900,878 | 2,582 | 46,039 | 16,916 | 56 | 12,229 | 7.67% | 13.49% |
| Yelp | 5,583,947 | 810,015 | 6,091 | 15,437 | 5,183 | 91 | 132,895 | 0.05% | 0.06% |
| Rees46 | 67,575,203 | 10,190,006 | 60 | 1,054,210 | 166,841 | 1 | 172,756 | 0.03% | 0.04% |
| Coveo | 1,411,113 | 165,673 | 17 | 52,501 | 7,748 | 1 | 10,868 | 0.00% | 0.00% |
| RetailRocket | 750,832 | 196,234 | 131 | 29,148 | 8,036 | 7 | 36,824 | 0.05% | 0.05% |

Dataset-Task Mismatch

# Resultados

- GRU v/s Feed Forward

Table 2. Recommendation accuracy using the same model with and without sequence modelling

| Dataset | Model w/ sequence modelling | | | | Model w/o sequence modelling | | | | Relative change | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Recall@N | | MRR@N | | Recall@N | | MRR@N | | Recall@N | | MRR@N | |
| | N=5 | N=20 | N=5 | N=20 | N=5 | N=20 | N=5 | N=20 | N=5 | N=20 | N=5 | N=20 |
| Rees46 | 0.3010 | 0.5293 | 0.1778 | 0.2008 | 0.2594 | 0.4785 | 0.1474 | 0.1694 | -13.80% | -9.58% | -17.09% | -15.67% |
| Coveo | 0.1496 | 0.3135 | 0.0852 | 0.1010 | 0.1289 | 0.2678 | 0.0734 | 0.0868 | -13.83% | -14.59% | -13.85% | -14.05% |
| Retailrocket | 0.3237 | 0.5186 | 0.1977 | 0.2175 | 0.2747 | 0.4652 | 0.1613 | 0.1806 | -15.13% | -10.30% | -18.42% | -16.97% |
| Amazon (Beauty) | 0.0784 | 0.1319 | 0.0527 | 0.0579 | 0.0779 | 0.1271 | 0.0531 | 0.0579 | -0.71% | -3.61% | 0.86% | 0.00% |
| MovieLens10M | 0.1728 | 0.3264 | 0.1062 | 0.1211 | 0.1276 | 0.2440 | 0.0763 | 0.0875 | -26.18% | -25.23% | -28.16% | -27.68% |
| Steam | 0.1117 | 0.2371 | 0.0662 | 0.0781 | 0.1035 | 0.2208 | 0.0622 | 0.0735 | -7.38% | -6.87% | -5.99% | -5.96% |
| Yelp | 0.0702 | 0.1627 | 0.0371 | 0.0457 | 0.0657 | 0.1625 | 0.0353 | 0.0445 | -6.46% | -0.12% | -4.78% | -2.51% |

# Overzealous preprocessing

Overzealous preprocessing

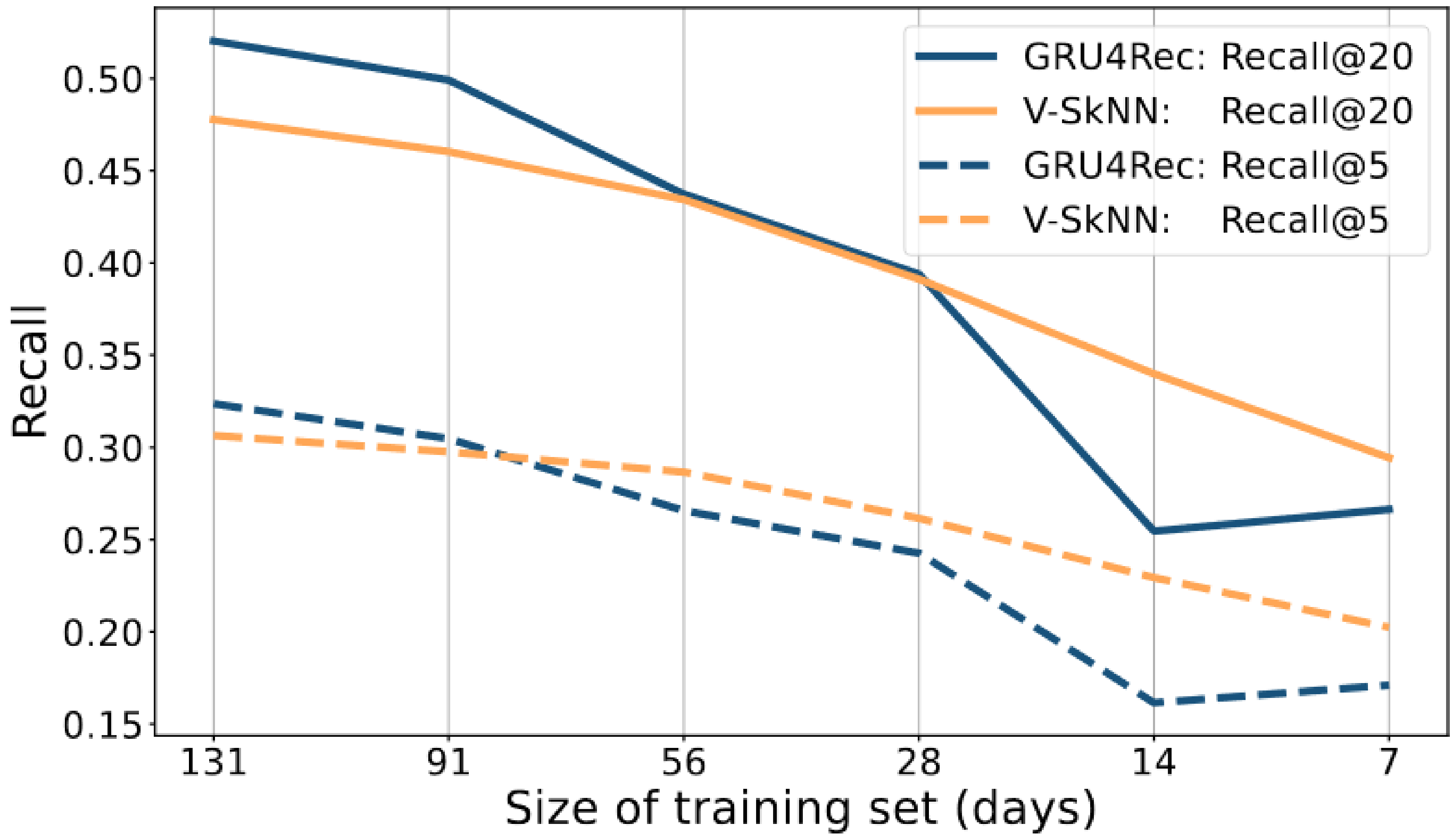
# Efectos del Preprocesamiento

- Los datos suelen tener ruido
- Considerar efectos del preprocesamiento
- La evaluación offline está sesgada
- Entre más fuerte sea el filtrado menos generales son las afirmaciones

Overzealous preprocessing

# Data training

- El tamaño del set de entrenamiento afecta el rendimiento del modelo

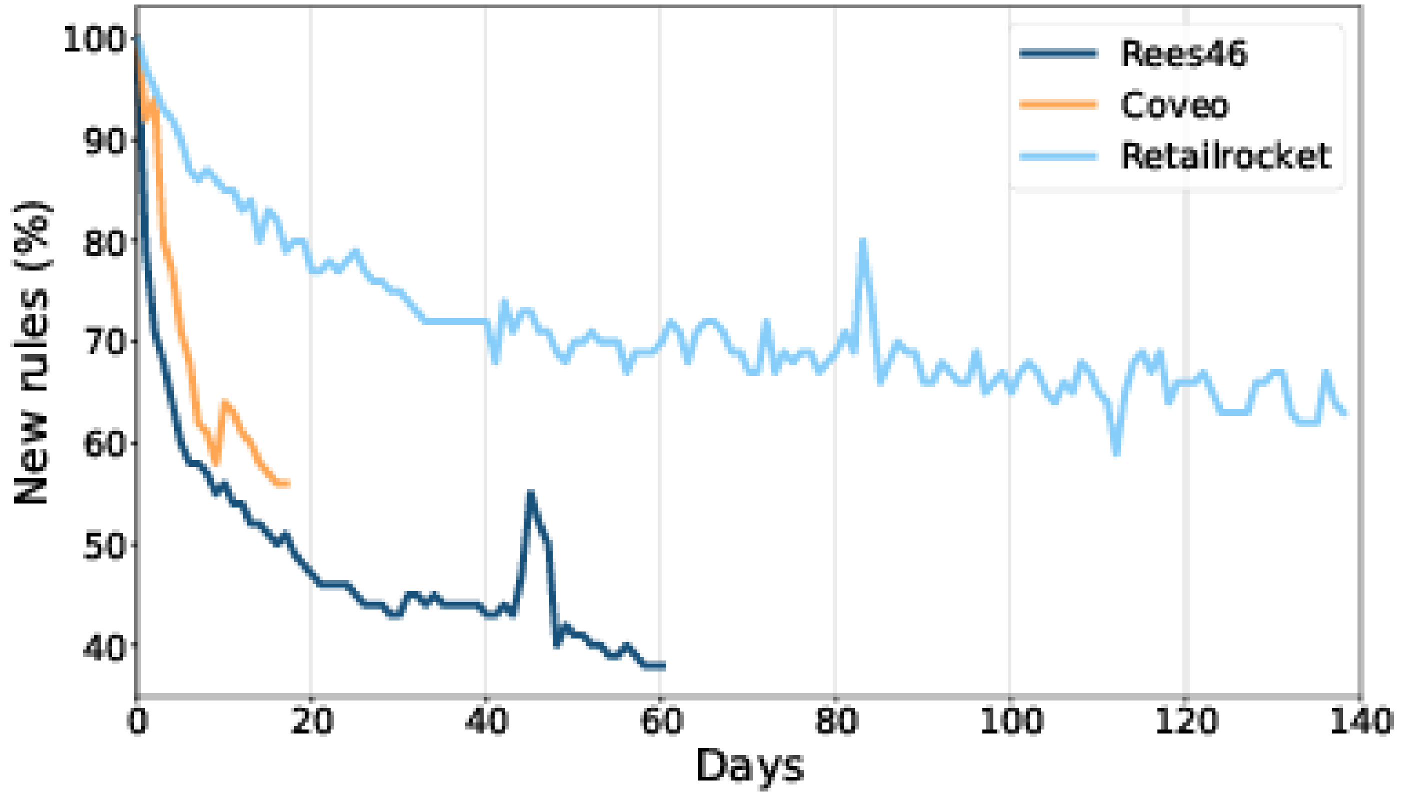


# Information leaking through time

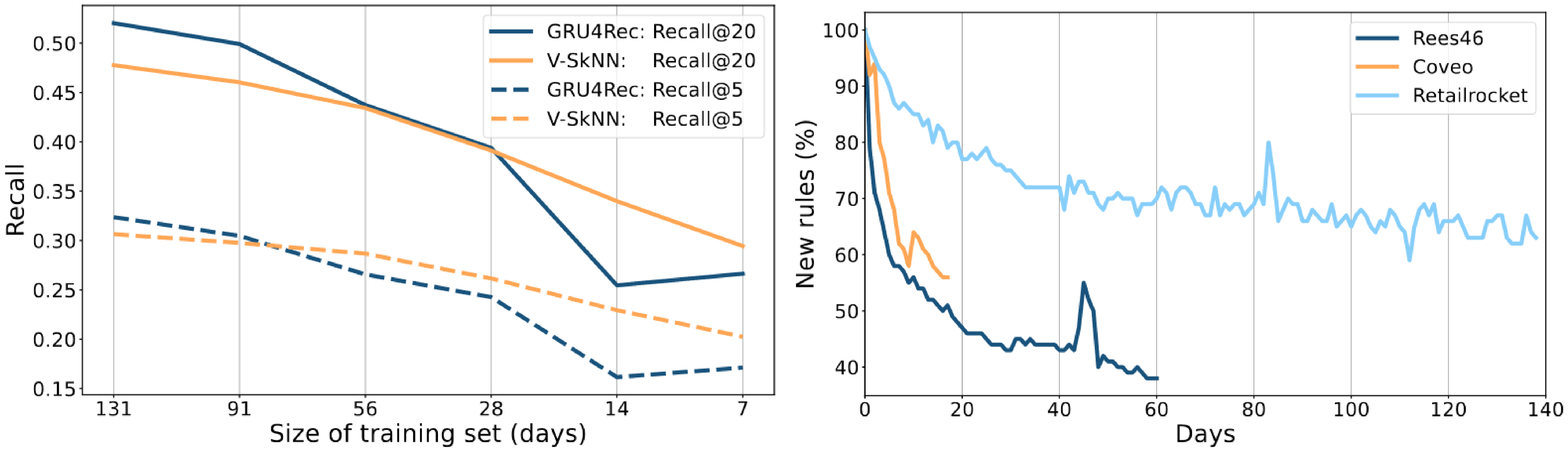
Information leaking through time

# Data training

- El tamaño del set de entrenamiento afecta el rendimiento del modelo

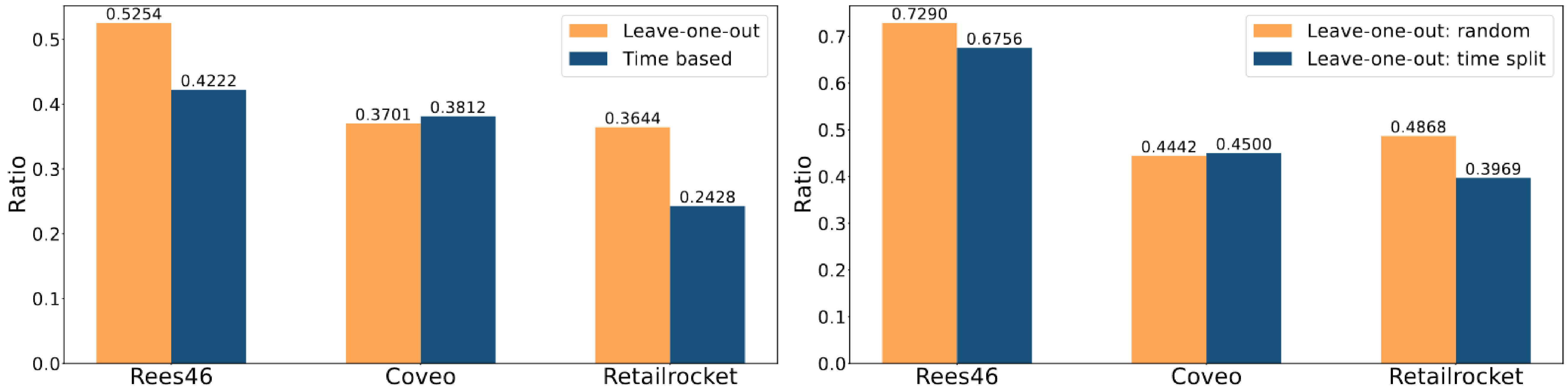



(b) Proportion of $i \to j$ item transitions observed first on day $N$ to the number of unique sequences of the same day




(a) The effect of using only recent data on the recommendation accuracy of model and neighbor based methods

(b) Proportion of $i \rightarrow j$ item transitions observed first on day $N$ to the number of unique sequences of the same day




(a) Leave-one-out and time based split

(b) Leave-one-out on random vs. most recent sessions

Fig. 2. Proportion of the $i \longrightarrow j$ test item transitions that are shared with the training set

# Negative sampling during testing

## Negative sampling during testing

# Razón de uso

- Conectado al cambio de error metrics a IR metrics
- Sampling en set de testeo
- Utilizado ampliamente

# Efectos en testing

- Sobreestimacion de metricas de evaluacion
- Cambia el ordenamiento de los modelos basado en el rendimiento

## Negative sampling during testing

# Resultados

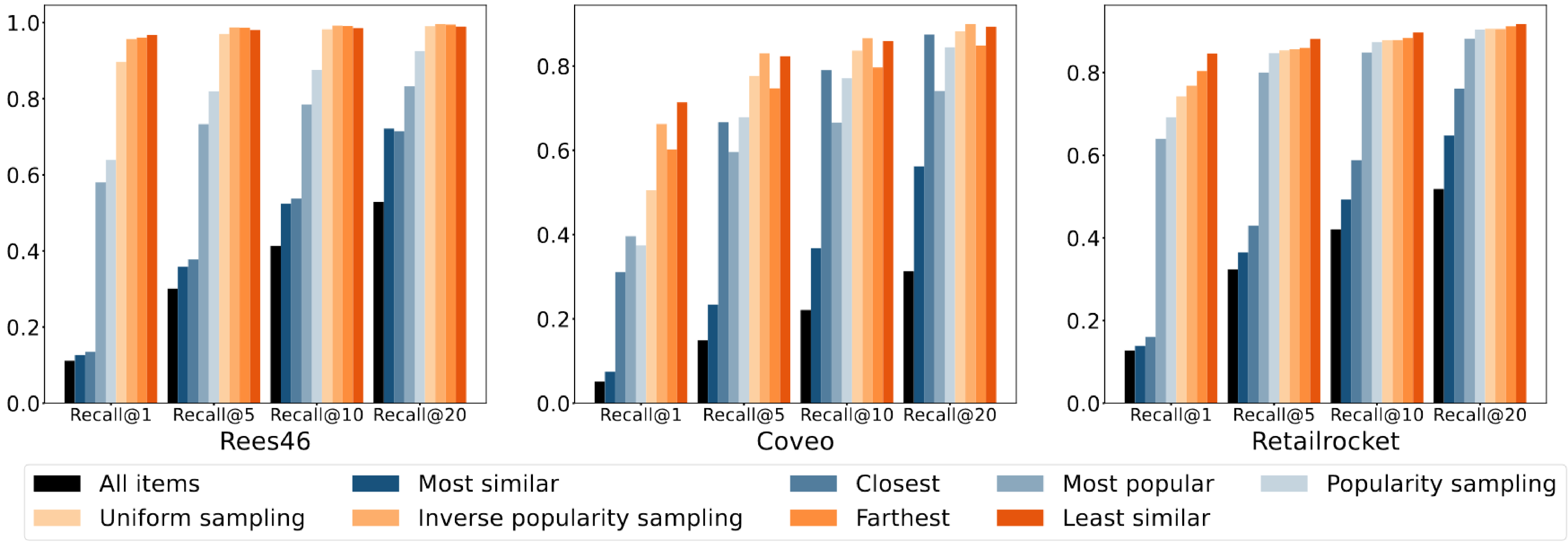


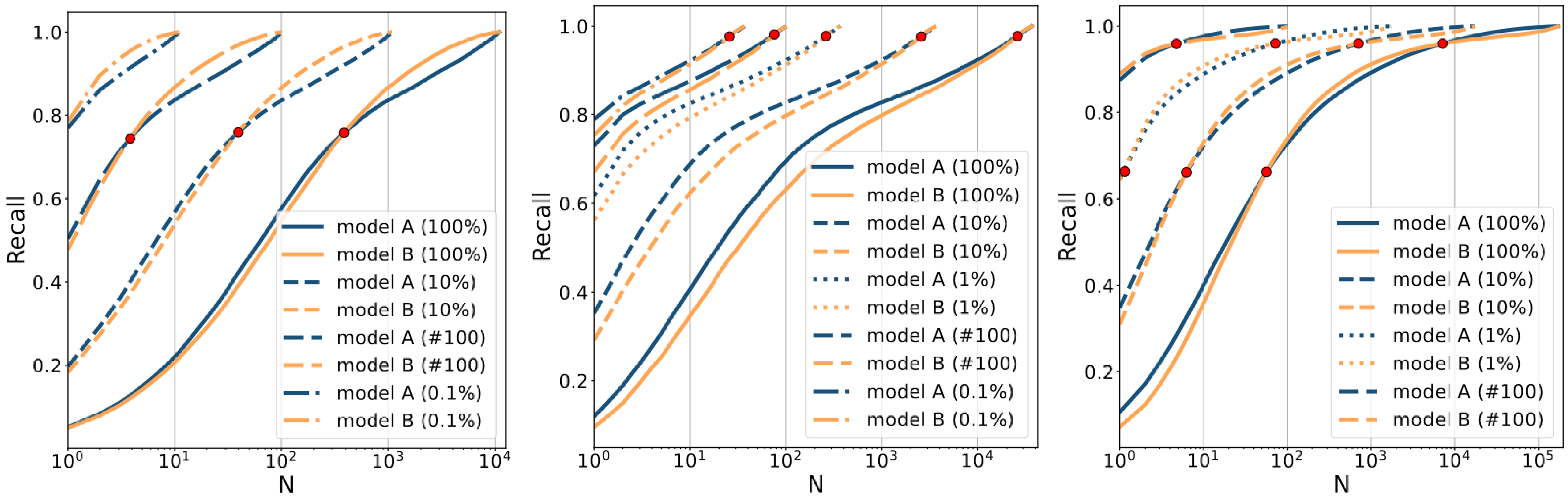

Fig. 3. Comparison of the strength of various negative samples of 100 items and no sampling.

# Resultados

- Disminucion de elementos negativos desafiantes
- Rendimiento relativo con longitud de lista de recomendación $M$ se desplaza a la longitud $N \, (\ll M)$

## Negative sampling during testing

# Resultados

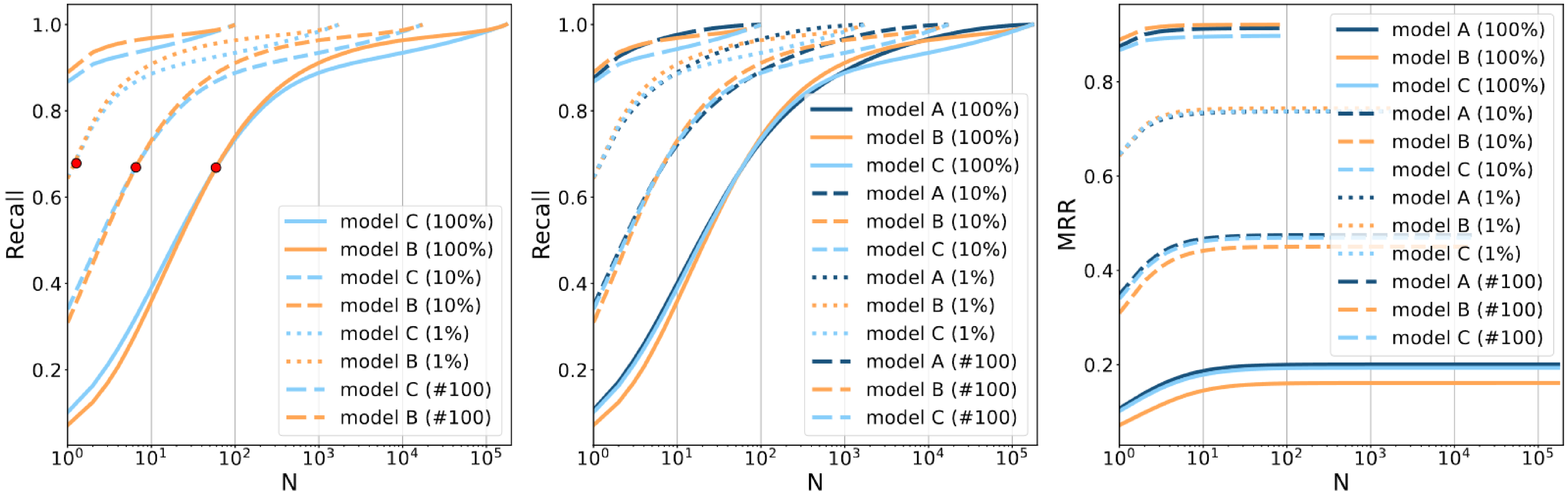



(a) Coveo – Recall@N (b) Retailrocket – Recall@N (c) Rees46 – Recall@N – AB

Fig. 4. Accuracy as the function of recommendation list length, with and without sampling

## Negative sampling during testing

# Resultados




(d) Rees46 – Recall@N – BC (e) Rees46 – Recall@N – ALL (f) Rees46 – MRR@N – ALL

Fig. 4. Accuracy as the function of recommendation list length, with and without sampling

# Trabajos Relacionados

# Trabajos relacionados

- La evaluacion offline ha sido discutida ampliamente
- Razon de falta de reproducibilidad actual (Ferrari, Cremonesi & Jannach, 2019 )
- Solo negative sampling ha sido estudiado (Krichene & Rendle, 2020)
- Breve mención de Dataset-Task Mismatch (Tang & Wang, 2018)

# Conclusiones

# Conclusiones

Overzealous preprocessing

Dataset-Task Mismatch

Negative sampling during testing

Information leaking through time

# Conclusiones

Dataset-Task Mismatch

Errores que plagan a la gran mayoría de estudios

Overzealous preprocessing

Negative sampling during testing

Information leaking through time

# Conclusiones

- Por tanto, antes de establecer las métricas y métodos de evaluación de un sistema recomendador, es necesario considerar todos los problemas presentados

Conclusiones

# Referencias


- Balázs Hidasi and Ádám Tibor Czapp. 2023. Widespread Flaws in Offline Evaluation of Recommender Systems. In Seventeenth ACM Conference on Recommender Systems (RecSys ’23), September 18–22, 2023, Singapore, Singapore. ACM, New York, NY, USA, 11 pages.
- Aleksandr Petrovand Craig Macdonald. 2022. A Systematic Review and Replicability Study of BERT4Rec for Sequential Recommendation. In Proceedings of the 16th ACM Conference on Recommender Systems. 436–447.
- Jin Huang, Wayne Xin Zhao, Hongjian Dou, Ji-Rong Wen, and Edward Y Chang. 2018. Improving sequential recommendation with knowledge-enhanced memory networks. In The 41st international ACM SIGIR conference on research & development in information retrieval. 505–514.
- Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In 2018 IEEE international conference on data mining(ICDM). IEEE, 197–206
- Wei Cai, Weike Pan, Jingwen Mao, Zhechao Yu, and Congfu Xu. 2022. Aspect Re-distribution for Learning Better Item Embeddings in Sequential Recommendation. In Proceedings of the 16th ACM Conference on Recommender Systems. 49–58

Conclusiones

# Referencias


- Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In Proceedings of the 26th international conference on world wide web. 173–182.
- Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. 2019. Are we really making much progress? A worrying analysis of recent neural recommendation approaches. In Proceedings of the 13th ACM conference on recommender systems. 101–109.
- Walid Krichene and Steffen Rendle. 2020. On sampled metrics for item recommendation. In Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining. 1748–1757.
- Jiaxi Tang and Ke Wang. 2018. Personalized top-n sequential recommendation via convolutional sequence embedding. In Proceedings of the eleventh ACM international conference on web search and data mining. 565–573.