Predictability Limits in Session-based Next Item Recommendation

Priit Järv
Tallinn University of Technology, Estonia
priit@whitedb.org

ABSTRACT

Session-based recommendations are based on the user's recent actions, for example, the items they have viewed during the current browsing session or the sightseeing places they have just visited. Closely related is sequence-aware recommendation, where the choice of the next item should follow from the sequence of previous actions.

We study seven benchmarks for session-based recommendation, covering retail, music and news domains to investigate how accurately user behavior can be predicted from the session histories. We measure the entropy rate of the data and estimate the limit of predictability to be between 44% and 73% in the included datasets.

We establish some algorithm-specific limits on prediction accuracy for Markov chains, association rules and k-nearest neighbors methods. With most of the analyzed methods, the algorithm design limits their performance with sparse training data. The session based k-nearest neighbors are least restricted in comparison and have room for improvement across all of the analyzed datasets.

CCS CONCEPTS

• **Information systems** → **Recommender systems**; *Test collections*; • **Mathematics of computing** → *Information theory*.

KEYWORDS

Session-based recommendation; Predictability

ACM Reference Format:
Priit Järv. 2019. Predictability Limits in Session-based Next Item Recommendation. In *Thirteenth ACM Conference on Recommender Systems (RecSys '19), September 16–20, 2019, Copenhagen, Denmark.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3298689.3346990

1 INTRODUCTION

Recommender systems aim to improve the user experience in shops, entertainment or travel apps and targeted advertising, by finding items or activities that the user is interested in. From the business point of view, the user is also more likely to make purchases when the recommender is able to anticipate what the user likes. However, the preferences of the user may not be readily available. The user may interact with the system anonymously or have no previous history. The user may be looking for place recommendations in a country or area they have not previously been to.

From practical perspective, being able to make recommendations in such scenarios is highly relevant. The problem of making recommendations tailored to user sessions, rather than established users, is called session-based recommendation. For the remainder of the paper, we will refer to any sequence of user actions as a *session* and the elements of the sequence as *items*. The items may refer to listened songs, visited sightseeing places or actual items in an online store.

In a related scenario, recommendations may themselves consist of a collection or a sequence of items. Music playlists and trip itineraries are examples where both the individual items and transitions between them are important for the quality of recommendations. Even when the goal is to recommend a single next item, such as the next song to listen, the sequence of previous items is often relevant. Recommender systems that consider the sequence of items are called sequence-aware.

The ability to predict the next item in a sequence correctly is commonly used to evaluate algorithmic approaches in session-based scenarios. While the quality in recommendations goes beyond just being able to find the most likely item, this is still an important indicator in determining whether the model has captured user interests accurately.

Differences in methodology, baselines and evaluation datasets in session-based recommendation research have made comparison and analysis of different algorithmic approaches difficult. To establish a common baseline for evaluating next item predictions, Ludewig and Jannach published a set of benchmark datasets and a standardized methodology [5]. They measured the performance of various methods and concluded that there are no major differences between simple methods like k-nearest neighbors and sophisticated methods like matrix factorization approaches that are intended to overcome the difficulties with simpler methods, such as data sparsity and ability to generalize. In fact, the simpler methods performed better on majority of the datasets.

In this paper, we examine properties of seven public datasets included in the benchmarks of Ludewig and Jannach [5]. To understand what kind of expectations could be placed on the accuracy of future improvements in algorithms, we estimate the predictability of the item sequences in the datasets. We define *predictability* as the probability that the recommender will correctly predict the next item, given an unfinished session and a history of other sessions. Assuming we knew the most likely item each time, the prediction can still be wrong when other items have a non-zero probability of appearing. When discussing the predictability of a dataset, we mean the probability that a perfect predictor will guess correctly, averaged over the dataset.

RecSys '19, September 16–20, 2019, Copenhagen, Denmark

ACM ISBN 978-1-4503-6243-6/19/09…$15.00
https://doi.org/10.1145/3298689.3346990

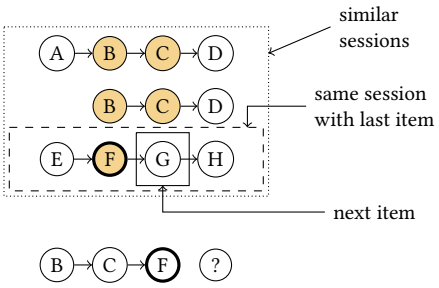

Figure 1: Items available for recommendation. Given a session to recommend for, the algorithms we analyze restrict the possible responses to co-occurrences in the training set with the session items {B, C, F} (marked with color fill).

For simpler algorithms, the design of the algorithm places an upper limit on the ability to predict the next item. For example, when modeling the sessions as a Markov chain, the probability distribution of the next item can only include the items that immediately followed the current item in a session in the training data. The method's accuracy is then limited to how often the known item pairs occur in the testing data, relative to previously unknown item pairings.

Figure 1 illustrates typical cases where the co-occurrence in the training data determines which items are available for recommendation. For a session {B, C, F}, considering only the continuations seen in training data leads to recommending G. Methods that look at entire sessions could also rank E and H. Neighborhood-based methods that find similar sessions can also recommend the item D.

We determine co-occurrence relations that affect the performance of Markov chains, association rules and k-nearest neighbors (k-NN) in all included datasets. These relations are highly dependent on datasets and training-testing splits, but are useful in explaining and further evaluating the performance of the affected methods in the reference benchmark of Ludewig and Jannach [5].

The remainder of the paper is organized as follows. In Section 2 we review the related work. In Section 3 we estimate two types of limits on prediction accuracy - upper limits due to the inherent randomness in the benchmark datasets and limits due to algorithm design. Section 4 summarizes the findings and suggests directions for future research.

2 RELATED WORK

This paper relies extensively on the evaluation of session-based recommenders by Ludewig and Jannach [5]. Recognizing the lack of unified methodology, they collected datasets and performed evaluation of multiple classes of methods. The public datasets and testing framework have also been published online, to facilitate building future research on this platform. They found that with their methodology and an increased variety of datasets there was no significant difference between naive baselines and latest sophisticated approaches. Ludewig and Jannach concluded that more research is required into the circumstances that affect algorithm performance in different datasets. We have made an initial investigation into this question in this paper.

Session-based recommendation can be seen as a subset of sequence-aware recommendation. Quadrana et al. review the applications, types and methods of sequence-aware recommenders [7]. In recent years, the attention has turned to deep learning approaches to recommendation, reviewed by Zhang et al. [19].

The entropy method to estimate predictability of sequential information was popularized by Song et al. They used mobile phone usage logs to calculate the predictability of trajectories of individual mobile phone users [10]. An explanation of the theoretical basis and additional proofs were given by Smith et al. [9]. The details of the entropy estimation of the original method of Song et al. can interpreted in different ways. To address that Xu et al. published a clarified method with experimental evaluation [15].

Substantial prior work exists on making next-item predictions using the datasets included in our evaluation [2–4, 11, 12, 14, 16, 17]. These papers generally use different methodology so we do not compare their results directly to our estimated predictability. They also cover a wider range of recommendation scenarios, including established user histories and leveraging content information [2, 4, 12, 14, 16, 17].

3 LIMITS ON PREDICTION ACCURACY

In this section we determine the limits on prediction accuracy due to inherent randomness in data and due to algorithm design. We consider only the limits on the hit rate (HR) metric, which can be directly estimated with our approach. The hit rate metric measures whether the item to predict was included in the top-n recommended items (abbreviated as HR@n). We cover 7 public datasets included in the benchmark of Ludewig and Jannach[5]:

- *RSC15* - item views and purchases in online retail, published for the ACM RecSys 2015 Challenge[1];
- *TMALL* - online purchasing history from Tmall.com;
- *RETAILR* - user browsing histories from Retail Rocket[8];
- *AOTM* - music playlists from the Art of the Mix platform[6];
- *30MUSIC* - listening histories from last.fm[13];
- *NOWPLAYING* - "currently listening" tweets[18];
- *CLEF* - news article reads by users, a subset of the data used in the 2017 CLEF NewsREEL challenge[5].

We calculate the estimates on training and test splits created with the same methodology and settings as used by Ludewig and Jannach[5] and compare our limit estimations to the performance of algorithms as measured in their benchmark. The programs to reproduce the calculations and instructions to access the data are available online[1].

3.1 Limit on predictability

We use the method of Song et al. [10] to estimate the upper bound on predictability of sequential data through estimating the entropy rate on the sequences. To determine predictability in the benchmark setting, we treat the training and test splits as a single joint sequence. This simplification is safe as long as we do not consider the effects of sessions overlapping or having taken place in parallel important.

[1] https://github.com/priitj/recsys19

Let the next item in a sequence be X_i and the items preceding it h_i. The entropy associated with the next item X_i, as a measurement of how predictable its possible values are, is:

$$H(X_i|h_i) = -\sum_x P(X_i = x|h_i) \log_2 P(X_i = x|h_i) \quad (1)$$

The probability distribution $P(X_i|h_i)$ is the true probability distribution of the next item X_i. While we do not know the true distributions and cannot calculate the quantity $H(X_i|h_i)$ directly, it will form the theoretical basis of determining the maximum predictability of the sequence. The entropy rate of a sequence is[9]:

$$\mathcal{H}(\mathcal{X}) = \lim_{t\to\infty} \frac{1}{t} \sum H(X_t|h_t) \quad (2)$$

Given the entropy rate, the bound on maximum predictability $\Pi^{\overline{max}}$ can be found by numerically solving

$$\mathcal{H}(\mathcal{X}) = -\Pi^{\overline{max}} \log_2 \Pi^{\overline{max}} - (1 - \Pi^{\overline{max}}) \log_2 \frac{1 - \Pi^{\overline{max}}}{m - 1} \quad (3)$$

where m is the number of unique items. The full derivation of Eq. 3 is given by Smith et al. [9]. Finally, we substitute the theoretical $\mathcal{H}(\mathcal{X})$ with the estimate $S \approx \mathcal{H}(\mathcal{X})$ over a sequence of length n:

$$S = \frac{1}{\frac{1}{n}\sum_i \Lambda_i} \log_2 n \quad (4)$$

This is the corrected estimate by Xu et al.[15], where $\Lambda_i = k^{(i)}_{max} + 1$ and $k^{(i)}_{max}$ is defined as the length of longest sub-sequence starting from position i that appears as a continuous sub-sequence between positions $1 \ldots i-1$. Because we use an estimate of the entropy rate, the calculated predictability should be treated as an approximation. Furthermore, the experiments of Xu et al. show that while the corrected method we use is more accurate, it can also underestimate predictability [15].

We include training and test sets because we are interested in determining the entropy rate of the same stochastic process that "produced" the sessions in both sets. The sets are represented by arranging the sessions sequentially and introducing a session end marker e, which is placed between individual sessions. The appearance of the marker in sequence is then the event that the session ends, given all the previous sessions and the content of the current session, with probability $P(X_i = e|h_i)$. Similarly, the probability that the next session starts with item x, given all the previous sessions, is $P(X_i = x|e, h_{i-1})$. Therefore, the sequence can still be viewed as events produced by a stochastic process and Equations 1–2 apply.

We report the average $\Pi^{\overline{max}}$ over five evaluation splits for each dataset in Table 1. The estimated predictability $\Pi^{\overline{max}}$ is a limit on the HR@1 accuracy metric, the ratio of tests where the recommender is able to recommend the correct item as first in a ranked list. The limit is between 44% and 73%, depending on the dataset. Purely session-based recommendation algorithms should not be expected to improve above these values. The state of the art results (column *best known*) are significantly below the limit.

The practical implications of this result depend on the application. In many cases, offering more than one recommendation is useful, so HR@1 is a too strict measure and performance with $n > 1$ is more informative. On the other hand, where the exact prediction or the top position in a ranked list of recommendations is important, the performance will be directly bound by these limits. Finally, the tools we have developed allow calculating predictability estimates on other datasets.

Table 1: The estimate of predictability from entropy. $\Pi^{\overline{max}}$ is the upper limit on the HR@1 metric. *Best known* – HR@1 benchmark performance of the best algorithm on the same data [5].

	$\Pi^{\overline{max}}$	best known[5] HR@1
RSC15	0.65	0.18
TMALL	0.58	0.13
RETAILR	0.59	0.27
AOTM	0.44	0.0096
30MUSIC	0.73	0.20
NOWPLAYING	0.71	0.076
CLEF	0.64	0.12

3.2 Limits due to algorithm design

In this section we evaluate the algorithmic limits of the following methods:

- *MC* - Markov Chain, where the transition probabilities are direct statistical probabilities learned from item-to-item transitions in training data;
- *SR* - Sequential Rules: like Markov Chain, but items indirectly following the current item are included as possible transitions with a decaying weight;
- *AR* - Association Rules: transition probabilities are learned statistically by counting item co-occurrences in training sessions, regardless of sequence;
- *IKNN* - Item-based k-NN, where nearest neighbors are found by comparing item vectors;
- *SKNN* - A family of session-based k-NN, where the common feature is that the nearest neighbors are found by comparing session vectors.

We use the method definitions of Ludewig and Jannach. Further details on the specific implementation choices can be found in their paper [5]. We do not include matrix factorization and neural network methods. While similar analysis of their performance would be highly relevant, it would require a more sophisticated mathematical approach than we have taken here.

With several of the analyzed methods, direct statistical learning from the training data is applied. For example, if the training set included sessions $\{A, B\}$, $\{B, A\}$ and $\{C, A\}$ then with the association rule method we would predict the item B after A with the probability $\frac{2}{3}$ and item C with probability $\frac{1}{3}$. The method would not be able to predict any other items as it has no grounds to do so based on the training data.

With simple recommendation approaches we can directly determine the limit of prediction accuracy since the algorithm design itself restricts which items can be recommended. By determining how the algorithm finds the items to recommend, we count the test instances where it is impossible to give the correct prediction

because the required association has not occurred in the training data. The ratio of such cases gives us an exact upper bound on the HR@n metric. The limit applies for any n. We give the results by algorithm, averaged over five training and test splits, in Table 2.

These results apply to the testing scenario where the recommender is given an incomplete session and has to predict the next item in the session correctly. In the following analysis, we refer to the incomplete session as the *current session*, and the last known item as the *current item*. The item that the recommender is expected to find, is referred to as the *next item*.

Figure 2 shows the limits of prediction accuracy for the most restricted algorithms, compared to their HR@20 performance and the best HR@20 result (state of the art), as measured by [5]. For the MC method, the limit is the ratio of test cases where the next item occurred directly after the current item in training data. The SR and AR methods can recommend the next item if there is a session in training data where the next item occurs somewhere after the current item, or anywhere in the session, respectively.

Overall, Figure 2 suggests that these three algorithms have similar dataset-dependent behaviour. The limits are higher in the *RSC15* and especially the *CLEF* dataset. On the e-commerce and music datasets they are low enough to directly influence the performance. On the *RETAILR* dataset, all three algorithms have a hard limit on the HR@n performance that has already been exceeded by the best known benchmark result at HR@20. With SR, this extends to the *AOTM* and *30MUSIC* datasets and with MC to all e-commerce and music datasets.

The significance of the limits decreases when more training data is available. The *CLEF* and *RSC15* datasets have on the average 2200 and 48 sessions per unique item in the training splits used, making them relatively least sparse in terms of training samples per item. The music datasets are the most sparse, having 0.36 or fewer sessions per unique item.

k-NN methods fall under two different categories. The IKNN method finds the items to recommend by calculating the cosine similarity of binary item vectors where the elements correspond to sessions. If the next item has not occurred in the same session with the current item, the dot product of the vectors will be 0. The IKNN method therefore has the same hard limit on the performance in HR@n metrics as AR.

In contrast with the other analyzed methods, the family of SKNN methods is much less restricted by design. They find the neighborhood by comparing session vectors, where the elements correspond to items. Whenever the next item has co-occurred with any of the items in the current session, it can potentially be recommended. In the datasets we analyzed, this covers over 80% of test cases each time and in particular, 92% for *AOTM* where other approaches are severely restricted.

Table 2: Limits on prediction accuracy (HR@n) for simpler approaches.

	MC	SR	AR, IKNN	SKNN
RSC15	0.79	0.86	0.90	0.97
TMALL	0.25	0.40	0.45	0.90
RETAILR	0.36	0.44	0.57	0.80
AOTM	0.015	0.042	0.079	0.92
30MUSIC	0.29	0.36	0.48	0.91
NOWPLAYING	0.17	0.28	0.40	0.90
CLEF	0.96	0.98	0.99	0.995

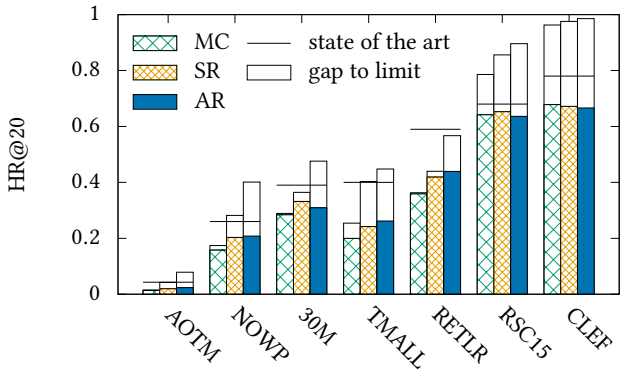

Figure 2: Performance (*HR@20* metric, colored bars) and remaining gap to limit due to algorithm design (clear bars). *State of the art* is the top result for the dataset (any algorithm). The performance measurements were taken from [5].

4 CONCLUSIONS AND FUTURE WORK

We gave estimates between 44% and 73% on the predictability of session-based recommendation scenarios with the analyzed datasets. The accuracy of the exact prediction of the next user action is limited by these values.

Algorithm-specific limits have a practical effect on the performance of the Markov chain, sequential rule, association rule and IKNN methods. While these methods were found to perform well in an earlier study [5], in several cases they cannot be improved significantly due to self-imposed restrictions. With sparse training data, their maximum theoretical performance is already below of what has been demonstrated in practice by other algorithms in the analyzed benchmark.

Session-based k-NN methods have room for improvement in all of the evaluated datasets. As they are already competitive with the more sophisticated approaches, more effort should be devoted into developing the SKNN algorithm family.

The methods presented in this paper have several limitations that should be addressed in future work. The estimates of predictability for other metrics that consider multiple recommendations and ranking positions, like the mean reciprocal rank (MRR) would be of practical value. The algorithm-specific analysis should be extended to important sequential recommendation methods – matrix factorization and neural networks.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous referees for their valuable comments and helpful suggestions.

REFERENCES

[1] David Ben-Shimon, Alexander Tsikinovsky, Michael Friedmann, Bracha Shapira, Lior Rokach, and Johannes Hoerle. 2015. RecSys Challenge 2015 and the YOOCHOOSE Dataset. In *Proceedings of the 9th ACM Conference on Recommender Systems, RecSys 2015, Vienna, Austria, September 16-20, 2015*, Hannes Werthner, Markus Zanker, Jennifer Golbeck, and Giovanni Semeraro (Eds.). ACM, 357–358. https://dl.acm.org/citation.cfm?id=2798723

[2] Negar Hariri, Bamshad Mobasher, and Robin D. Burke. 2012. Context-aware music recommendation based on latenttopic sequential patterns. In *Sixth ACM Conference on Recommender Systems, RecSys '12, Dublin, Ireland, September 9-13, 2012*, Padraig Cunningham, Neil J. Hurley, Ido Guy, and Sarabjot Singh Anand (Eds.). ACM, 131–138. https://dl.acm.org/citation.cfm?id=2365979

[3] Balázs Hidasi and Alexandros Karatzoglou. 2018. Recurrent Neural Networks with Top-k Gains for Session-based Recommendations. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018*, Alfredo Cuzzocrea, James Allan, Norman W. Paton, Divesh Srivastava, Rakesh Agrawal, Andrei Z. Broder, Mohammed J. Zaki, K. Selçuk Candan, Alexandros Labrinidis, Assaf Schuster, and Haixun Wang (Eds.). ACM, 843–852. https://doi.org/10.1145/3269206.3271761

[4] Dietmar Jannach, Iman Kamehkhosh, and Lukas Lerche. 2017. Leveraging multi-dimensional user models for personalized next-track music recommendation. In *Proceedings of the Symposium on Applied Computing, SAC 2017, Marrakech, Morocco, April 3-7, 2017*, Ahmed Seffah, Birgit Penzenstadler, Carina Alves, and Xin Peng (Eds.). ACM, 1635–1642. https://doi.org/10.1145/3019612.3019756

[5] Malte Ludewig and Dietmar Jannach. 2018. Evaluation of session-based recommendation algorithms. *User Model. User-Adapt. Interact.* 28, 4-5 (2018), 331–390. https://doi.org/10.1007/s11257-018-9209-6

[6] Brian McFee and Gert R. G. Lanckriet. 2012. Hypergraph Models of Playlist Dialects. In *Proceedings of the 13th International Society for Music Information Retrieval Conference, ISMIR 2012, Mosteiro S.Bento Da Vitória, Porto, Portugal, October 8-12, 2012*, Fabien Gouyon, Perfecto Herrera, Luis Gustavo Martins, and Meinard Müller (Eds.). FEUP Edições, 343–348. http://ismir2012.ismir.net/event/papers/343-ismir-2012.pdf

[7] Massimo Quadrana, Paolo Cremonesi, and Dietmar Jannach. 2018. Sequence-Aware Recommender Systems. *ACM Comput. Surv.* 51, 4 (2018), 66:1–66:36. https://doi.org/10.1145/3190616

[8] Retailrocket. 2017. Retailrocket recommender system dataset. https://www.kaggle.com/retailrocket/ecommerce-dataset.

[9] Gavin Smith, Romain Wieser, James Goulding, and Duncan Barrack. 2014. A refined limit on the predictability of human mobility. In *IEEE International Conference on Pervasive Computing and Communications, PerCom 2014, Budapest, Hungary, March 24-28, 2014*. IEEE Computer Society, 88–94. https://doi.org/10.1109/PerCom.2014.6813948

[10] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. 2010. Limits of predictability in human mobility. *Science* 327, 5968 (2010), 1018–1021.

[11] Yong Kiam Tan, Xinxing Xu, and Yong Liu. 2016. Improved Recurrent Neural Networks for Session-based Recommendations. In *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems, DLRS@RecSys 2016, Boston, MA, USA, September 15, 2016*, Alexandros Karatzoglou, Balázs Hidasi, Domonkos Tikk, Oren Sar Shalom, Haggai Roitman, Bracha Shapira, and Lior Rokach (Eds.). ACM, 17–22. https://doi.org/10.1145/2988450.2988452

[12] Roberto Turrin, Andrea Condorelli, Paolo Cremonesi, Roberto Pagano, and Massimo Quadrana. 2015. Large scale music recommendation. In *Workshop on Large-Scale Recommender Systems (LSRS 2015) at ACM RecSys*.

[13] Roberto Turrin, Massimo Quadrana, Andrea Condorelli, Roberto Pagano, and Paolo Cremonesi. 2015. 30Music Listening and Playlists Dataset. In *Poster Proceedings of the 9th ACM Conference on Recommender Systems, RecSys 2015, Vienna, Austria, September 16, 2015. (CEUR Workshop Proceedings)*, Pablo Castells (Ed.), Vol. 1441. CEUR-WS.org. http://ceur-ws.org/Vol-1441/recsys2015_poster13.pdf

[14] Pengfei Wang, Jiafeng Guo, Yanyan Lan, Jun Xu, Shengxian Wan, and Xueqi Cheng. 2015. Learning Hierarchical Representation Model for NextBasket Recommendation. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, August 9-13, 2015*, Ricardo A. Baeza-Yates, Mounia Lalmas, Alistair Moffat, and Berthier A. Ribeiro-Neto (Eds.). ACM, 403–412. https://doi.org/10.1145/2766462.2767694

[15] Paiheng Xu, Likang Yin, Zhongtao Yue, and Tao Zhou. 2019. On predictability of time series. *Physica A: Statistical Mechanics and its Applications* 523 (2019), 345–351.

[16] Haochao Ying, Fuzhen Zhuang, Fuzheng Zhang, Yanchi Liu, Guandong Xu, Xing Xie, Hui Xiong, and Jian Wu. 2018. Sequential Recommender System based on Hierarchical Attention Networks. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden.*, Jérôme Lang (Ed.). ijcai.org, 3926–3932. https://doi.org/10.24963/ijcai.2018/546

[17] Feng Yu, Qiang Liu, Shu Wu, Liang Wang, and Tieniu Tan. 2016. A Dynamic Recurrent Model for Next Basket Recommendation. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016*, Raffaele Perego, Fabrizio Sebastiani, Javed A. Aslam, Ian Ruthven, and Justin Zobel (Eds.). ACM, 729–732. https://doi.org/10.1145/2911451.2914683

[18] Eva Zangerle, Martin Pichl, Wolfgang Gassler, and Günther Specht. 2014. #nowplaying Music Dataset: Extracting Listening Behavior from Twitter. In *Proceedings of the First International Workshop on Internet-Scale Multimedia Management, WISMM '14, Orlando, Florida, USA, November 7, 2014*, Roger Zimmermann and Yi Yu (Eds.). ACM, 21–26. https://doi.org/10.1145/2661714.2661719

[19] Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. 2019. Deep Learning Based Recommender System: A Survey and New Perspectives. *ACM Comput. Surv.* 52, 1 (2019), 5:1–5:38. https://dl.acm.org/citation.cfm?id=3285029