Characteristics of the Learning Data of a Session-Based Recommendation System and Their Impact on the Performance of the System

Urszula Kużelewska
Faculty of Computer Science/Bialystok University of Technology
Wiejska 45a 15-351 Bialystok, Poland *u.kuzelewska@pb.edu.pl*

Małgorzata Charytanowicz
Department of Computer Science/Lublin University of Technology
Nadbystrzycka 36B 20-618 Lublin, Poland *m.charytanowicz@pollub.pl*

Abstract

Recommendation systems are an effective solution for personalising e-commerce services. They are able to provide customers with relevant and useful products. Their performance is determined by the quality of the methods employed. However, it is also influenced by the input data. Session-based (SB) techniques are highly effective in real-world scenario to generating recommendations that focus on short-term user activities. This study aims to investigate the relation between data statistics and performance of SB algorithms measured by accuracy and coverage.

Keywords: Session-based recommenders, Evaluation of recommender systems, Data statistics.

1. Introduction

A recommender system efficiently retrieves relevant information from a vast amount of data. These applications function as digital advisors, collecting behavioural information on users and providing personalised recommendations [9], [14]. Session-based recommenders are a collaborative filtering (CF) approach, that predict the next action in the user's session, which time is limited to minutes. That means that no historical data is stored for individual users. Any user interactions that occur after the ongoing session has expired are treated as new ones [13].

Data can be described by different characteristics and may affect the performance of recommendations [10], [12]. In [11] an in-depth comparison of 12 SB methods on 8 datasets is presented. The study revealed that there is no single algorithm that consistently outperforms the others.

This work aimed to determine whether a relationship exists between data and a SB system's performance. The selected factors include data density, shape, and item popularity [7], [3]. The accuracy of the system was calculated by the HitRate index, and its diversity - by Coverage.

This paper is inspired by [1] and [4]. Nevertheless, the distinction is as follows. First, the comparison was between SB recommenders and the traditional CF approach. Then, the dataset in the experiments was Diginetica [11], an e-commerce dataset utilised in the RecSys Challenge. The set originally comprised 275,000 sessions and 160,000 items. The aforementioned studies were based on two MovieLens datasets, comprising 100,000 and 1,000,000 ratings, respectively. The nature of the data was also distinct, as sessions comprised records of users' activity, rather than ratings. Furthermore, the MovieLens dataset is dense, comprising users who have rated at least 20 movies. As with the other datasets employed in SB systems, Diginetica is notably sparse. The performance measures differed as well. In [1], were analysed error-based metrics, while in [4], the authors also considered fairness. This paper examines another quality and diversity metric, namely Hit Rate and Coverage.

This work provides the following contribution. The impact of input data characteristics on

the performance of session-based recommender systems is not equal across recommender types and should be analysed individually.

2. Related Work

The existing literature contains a number of works that address the subject of learning data correlation and the performance of recommender systems. Nevertheless, these studies merely indicate some potential relations and propositions for overcoming the issue without a comprehensive investigation of this topic.

According to Hsu's research [8], skewness is a characteristic that reduces the accuracy of CF methods. It was was confirmed through experiments conducted on a naturally skewed real dataset containing clickstreams from an advertising online agency. In [15] a novel approach to addressing data sparsity was applied, comprising data augmentation and refinement, with the objective of improving data characteristics and, in turn, the accuracy of recommendations.

There are only few recent studies that have extensively examined the effect of data characteristics on classical recommender systems' performance, providing valuable insights into the correlation between data characteristics and recommendation accuracy [4].

The objective of the work [6] was to examine the influence of data characteristics on the efficacy of the most prevalent shilling attacks against popular CF methods. The results provided sufficient statistical evidence to demonstrate that data characteristics, in particular, size, shape, and density, are important factors in determining the effectiveness of an attack. Furthermore, the study identified the most significant features with respect to a specific type of recommender.

The authors of [5] conducted a comprehensive literature review with the objective of examining the characteristics of the datasets utilized in traditional collaborative filtering recommender systems. The aim of this review was to identify similarities and differences between these datasets, with the ultimate goal of providing researchers with a set of guidelines to assist them in selecting appropriate datasets for their experiments. The following indices were investigated: Shape, Space, Density, and Gini. The findings demonstrated that datasets with markedly disparate characteristics enhance the robustness of the evaluation process.

The most recent and comprehensive work [4] proposed an explanatory framework based on regression models to better understand how data characteristics impact on the fairness and accuracy of recommender systems. The researchers considered a number of data characteristics, including those related to the structure of the rating matrix, or the rating frequency distribution. The results demonstrated that the three most significant characteristics may contribute up to 80–90% towards the overall accuracy of a recommender In the case of the systems' fairness, however, such a relationship is not evident.

3. A Dataset and its Characteristics

In SB systems, it is possible to create a User Rating Matrix (URM) where the values are binary and indicate whether the user is interested in the item. The URM is a matrix with columns and rows corresponding to the system's items (V) and users (U). The shape indicates the ratio of users to products in the system (1). Another crucial aspect is data density (see also (1)).

$$Shape(URM) = \frac{|U|}{|V|}, \quad Density(URM) = \frac{n_r}{|U| \cdot |V|} \tag{1}$$

where n_r is a sum of all session lengths in the matrix URM.

The impact of popular products on the efficiency of a recommendation algorithm is a significant factor [16]. The products often present in user sessions will be recommended to them more often, resulting in reduced system efficiency and less diversity in recommendations [4]. The example measures are the AvgPop and the long-tail skewness coefficient (LTS) (2).

$$AvgPop(URM) = \frac{1}{|U|} \cdot \frac{\sum_{k \in Ru} \phi(i)}{R_u}, \quad LTS(URM) = \frac{1}{|V|} \cdot \frac{\sum_{i=1}^{|V|} (\phi(i) - \mu)^3}{[\frac{1}{|V|} \sum_{k=1}^{|V|} (\phi(i) - \mu)^2]^{\frac{3}{2}}} \quad (2)$$

where R_u is a set of items in a session, μ is an average overall popularity of all items.

The AvgPop metric calculates the average popularity of items across sessions. An item's popularity score (denoted as $\phi(i)$) is determined by the number of users interacting with it across the entire user set. The LTS coefficient is more sensitive to the actual popularity with respect to the size of the long tail items.

The experiments described below were conducted on 51 subsets of Diginetica, prepared to obtain certain values of the characteristics [4]. First, the set was divided into 5 equal sets. Then, the subsets were generated from the 5 main sets, starting from the current statistic's value. Sessions or items were removed randomly to obtain diversified characteristics - see Table 1. Each test involved at least 25 subsets.

Table 1. Characteristics of the data from the subsets used in the experiments

	Actions	**Sessions**	**Items**	**Actions/Sessions**	**Actions/Items**
min	21786	12632	5000	1.29	1.04
max	165084	57377	35428	5.99	6.49
	Shape	**Density**	**AvgPop**	**LTS**	
min	0.75	$8.24 \cdot 10^4$	1.33	-14 153	
max	2.73	$34.49 \cdot 10^4$	17.84	-2002	

4. Experimental Setup and Results

For each experiment, a unique set of data was selected to consider various aspects. In the first and second experiment, which focus on the Shape and Density of the rating matrix, it was aimed to obtain data with varying ratios of users to products. By providing different numbers of users and items while keeping the other statistics constant, different Shape and Density rates were achieved. The following experiments were concentrated on various factors of popularity of items. The products were sorted by their popularity values and confidently iteratively removed to retain the same number of users. The resulting sets were analysed in terms of the skewness coefficient, and only sets with significant changes in the coefficient were selected.

Accuracy was quantified using the HitRate index, a standard metric employed in SB approaches. The procedure for calculating based on the evaluation of the content of recommendation lists when successive items are incrementally added to the test sessions. Then, after generating the propositions, the list is truncated at the particular position and the content is examined in terms of the presence of the items from the test vector. Commonly, short and long thresholds are examined, here: 3 and 20 elements on the recommendation lists (HR@3 and HR@20). Coverage [2] indicates the frequency with which items appear in the recommendation lists. Its low values relate to the tendency to recommend the same set to many users. In the majority of cases, the coverage cut-off is equal to 20.

Recommendations were obtained using session-based collaborative filtering algorithms (implementations from the Session-Rec library [17]). The following methods were used: SKNN (a neighbourhood-based approach), and STAMP (based on a neural network).

4.1. Obtained Results

The prepared datasets were used to generate and evaluate recommendation lists. Cross-validation was used with a minimum split of 27 sets and results averaging. The recommendation lists were

then evaluated according to the following indices: HR and Coverage, with a cut-off threshold of 3 or/and 20 elements on the recommendation lists. The results are presented in Table 2.

Table 2. Correlation of data characteristics and recommendation accuracy.

Name of charact.	**STAMP**			SKNN		
	HR@3	**HR@20**	**Coverage@20**	**HR@3**	**HR@20**	**Coverage@20**
Shape	0.09	0.23	-0.15	0.80	0.90	-0.07
Density	0.81	0.78	0.40	0.38	0.53	0.12
AvgPop	-0.28	-0.35	0.10	-0.73	-0.92	0.57
LTS	0.33	0.40	-0.08	0.87	0.97	-0.15

The first experiment focused on the shape of the rating matrix. It can be seen that the outcomes for both algorithms are different. For STAMP there are no significant positive changes in any of the metrics based on the shape of the data matrix. Whereas, for SKNN an increases in accuracy is observed as the Shape value grows. There is also a correlation between Shape and HitRate - the value is 0.80 (HR@3) and 0.90 (HR@20). In the case of the SKNN algorithm, the correlation is not particularly strong: 0.09 (HR@3) and 0.23 (HR@20). The Shape index was not found to be significantly correlated with Coverage.

The evaluation results for the Density index confirm a significant relationship between data density and the accuracy of the recommendation lists. The HitRate increases with increasing data density. However, the strength varies. For STAMP, it is definitely higher (0.81 and 0.78 for HR@3 and HR@20 respectively). For SKNN the values are 0.33 and 0.62 respectively, which means that the relationship is weaker. A small correlation was identified in the case of STAMP with regard to Coverage (0.4).

The results for Average Popularity and LongTailSkewness are as follows. A high average popularity of all items has a negative impact on recommendation accuracy. However, the correlations between AvgPop values and HR@3 are as follows: -0.28 for STAMP and -0.73 for SKNN. The correlation between AvgPop and HR@20 is -0.35 for STAMP and -0.92 for SKNN. Consequently, the neighbourhood-based approach is more susceptible to the average popularity of items. The LongTailSkewness exerts a somewhat more pronounced influence on the accuracy of recommendations. The values for HR@3 and HR@20 are 0.87 and 0.97 (for SKNN) and 0.33 and 0.40, respectively, for HR@3 and HR@20 (for STAMP). In conclusion, STAMP exhibits a greater capacity for personalisation, as evidenced by its lower correlation values. A correlation of 0.57 was identified in the case of SKNN with regard to Coverage.

5. Conclusions and Future Work

This paper presents experimental findings on the correlation between 4 data characteristics: Shape, Density, Popularity, and LTS, and the session-based recommender systems. The results indicate that certain methods are more strongly correlated than others. The STAMP neural network-based recommender demonstrated robust resistance to data features related to popularity bias, whereas it highlighted a high sensitivity to its density. In contrast, the SKNN neighbourhood-based system exhibited an opposite behaviour towards generated propositions, with accuracy related to the data nature, in particular an average popularity index.

Obviously, thoroughly examining and preparing the data is crucial to improving the accuracy of recommender systems. However, in ensemble approaches and commercial applications, it is of the utmost importance to select the recommenders carefully according to the data nature. The identification of correlated features with the accuracy of recommendation lists allows for the preparation and deployment of appropriate data improvement procedures.

The results are preliminary and therefore further experimentation is required to develop the assumptions. Initially, further datasets must be examined and subjected to a further comparison

of their characteristics. Subsequently, further algorithms will be investigated with respect to their types, for example, neighbourhood-based and neural network approaches. Finally, further data statistics will be considered, including Gini, skewness, and kurtosis of long-tail items.

Acknowledgments

The work was supported by a grant from the Bialystok University of Technology WZ/WI-IIT/3/2023 and funded with resources for research by the Ministry of Science and Higher Education in Poland.

The research leading to these results has received funding from the commissioned task entitled "VIA CARPATIA Universities of Technology Network named after the President of the Republic of Poland Lech Kaczyński" contract no. MEiN/2022/DPI/2575 action entitled "In the neighborhood – inter-university research internships and study visits."

References

1. Adomavicius G., Zhang J.: Impact of data characteristics on recommender systems performance, ACM Transactions on Management Information Systems **3**(1), 1-17 (2012)
2. Adomavicius, G., Kwon, Y.O.: Improving aggregate recommendation diversity using ranking-based techniques. IEEE Trans. Knowl. Data Eng. **24** (5), 896–911 (2012)
3. Alharbey R. et al.: Modeling user rating pref. behavior to improve the performance of the collaborative filtering based recommender systems, PLOS ONE **14**(8), (2019)
4. Bellogin A., Deldjoo Y., Di Noia T.: Explaining recommender systems fairness and accuracy through the lens of data characteristics, Inform. Process.&Manag. **58** (5), (2021)
5. Chin, J.Y. et al.: The Datasets Dilemma: How Much Do We Really Know About Recommendation Datasets? In: 15th ACM Int. Conf. WSDM '22, pp 141–149 (2022)
6. Deldjoo, Y. et al.: How Dataset Characteristics Affect the Robustness of Collaborative Rec. Models. In: 43rd Int. ACM SIGIR Conf. SIGIR '20, pp. 951–960 (2020)
7. Gunawardana A., Shani G.: Evaluating Recommendation Systems, In: Recommender Systems Handbook, Springer (2011)
8. Hsu, C. N., Chung, H. H., and Huang, H. S.: Mining skewed and sparse transaction data for personalized shopping recommendation. Machine Learning **57**(1), pp. 35-59 (2004)
9. Jannach, D., Mobasher, B., Berkovsky, S.: Research directions in session-based and sequential recommendation. User Model.-User-Adap. Interaction **30**, 609–616 (2020)
10. Kużelewska, U.: Scheme Selection Based on Clusters' Quality in Multi-Clustering M-CCF Recommender System. In: ISD2023 Proceedings (2023)
11. Ludewig, M., Mauro, N., Latifi, S. et al.: Empirical analysis of session-based recommendation algorithms. User Model User-Adap Inter 31, 149–181 (2021)
12. Ozdemir, S., Susarla, D.: Feature Engineering Made Easy: Identify unique features from your dataset in order to build powerful machine learning systems. Packt (2018)
13. Quadrana, M., Cremonesi, P., Jannach, D.: Sequence-aware recommender systems. ACM Comput. Surv. 54, 1–36 (2018)
14. Ricci, F., Rokach, L., Shapira, B.: Recommender Systems: Introduction and Challenges. Recommender systems handbook, pp. 1–34. Springer (2015)
15. Shaikh, S. et al.: Data augmentation and refinement for recommender system: A semi-supervised approach using max. margin matrix fact., Expert Syst Appl **238** (B) (2024)
16. Smyth, B., McClave, P.: Similarity vs. diversity. In Proceedings of the International Conference on Case-Based Reasoning, Springer, pp. 347—361 (2001)
17. Session-rec software https://github.com/rn5l/session-rec. Accessed February 10, 2024.