A Systematical Evaluation for Next-Basket Recommendation Algorithms

Zhufeng Shao*, Shoujin Wang†, Qian Zhang*, Wenpeng Lu*(✉), Zhao Li‡, Xueping Peng§
* School of Computer, Qilu University of Technology (Shandong Academy of Sciences), Jinan, China
† The Data Science Institute, University of Technology Sydney, Sydney, Australia
‡ Shandong Evay Info Technology Co., Ltd., Jinan, China
Shandong Computer Science Center (National Supercomputer Center in Jinan), Jinan, China
§ Australian Artificial Intelligence Institute, University of Technology Sydney, Sydney, Australia
Email: zhufengshao7@gmail.com, shoujin.wang@uts.edu.au, qianzhang9706@gmail.com,
lwp@qlu.edu.cn, liz@sdas.org, xueping.peng@uts.edu.au

***Abstract*—Next basket recommender systems (NBRs) aim to recommend a user's next (shopping) basket of items via modeling the user's preferences towards items based on the user's purchase history, usually a sequence of historical baskets. Due to its wide applicability in the real-world E-commerce industry, the studies NBR have attracted increasing attention in recent years. NBRs have been widely studied and much progress has been achieved in this area with a variety of NBR approaches having been proposed. However, an important issue is that there is a lack of a systematic and unified evaluation over the various NBR approaches. Different studies often evaluate NBR approaches on different datasets, under different experimental settings, making it hard to fairly and effectively compare the performance of different NBR approaches. To bridge this gap, in this work, we conduct a systematical empirical study in NBR area. Specifically, we review the representative work in NBR and analyze their cons and pros. Then, we run the selected NBR algorithms on the same datasets, under the same experimental setting and evaluate their performances using the same measurements. This provides a unified framework to fairly compare different NBR approaches. We hope this study can provide a valuable reference for the future research in this vibrant area.**

***Index Terms*—next basket recommendation, recommender systems, evaluation, fair comparison**

I. Introduction

Recent years have witnessed the great success of Recommender Systems (RSs) in many different real-world applications, such as E-commerce, stream media and online retail industry [1]–[4]. RSs have become a fundamental tool for users to make right choices from massive and redundant contents, products and services in an effective and efficient way [5]. As one of the most commonly used and practical RSs, next-basket recommender systems (NBRs), as a sub-area of RSs, have attracted increasing attention in recent years.

Why next-basket recommender systems? (1). Next-basket recommender systems (NBRs) are one of the most applicable RSs in the real-world shopping scenarios. The reason is that users often purchase a (shopping) basket of items rather than one single item in a shopping visit. For instance, Bob usually purchases a basket of daily products in his weekly shopping event. Therefore, NBRs can naturally match the real-world shopping scenarios since they aim to recommend a basket of items which are carefully selected to match a user's current demand and preference. (2). Although a variety of approaches for next-item recommendation task have been developed [6]–[9], they aim to recommend the next item within the current basket only via modeling the intra-basket correlations over items. As a result, they cannot be employed for the next-basket recommendation task in which a basket of inter-correlated items are recommended via modeling the sequential dependencies over a sequence of baskets. Therefore, new theories and approaches are in demand for next-basket recommendation, making NBR a significant research topic. (3). NBRs predict the next basket of items which mostly interest a user by capturing the user's preference from her/his purchase history, namely a sequence of baskets purchased recently. As a result, both the user's long-term and short-term preferences can be well modeled for more accurate recommendation.

Given a user's historical transaction records, usually a sequence of shopping baskets, a NBR aims at predicting the next basket of items that a user would like to purchase by modeling the sequential dependencies over the sequence of baskets. A variety of existing studies on next-basket recommendation have emerged with great success. For example, a Markov Chain (MC) based approach (Rendle et al.) [10] has been proposed to capture low-order dependencies over baskets for next-basket recommendation, embedding-based approaches (Wan et al.; Wang et al.) [11], [12] have been developed to use distributed representation to predict the next basket, recurrent neural network (RNN) based approaches (Yu et al.; Hu et al.; Le et al.; Bai et al.; Qin et al.) [2], [13]–[16] were proposed to capture higher-order dependencies across baskets. In addition, intention-driven approaches (Wang et al.) [17], [18] have been proposed to model the heterogeneous intentions contained in the historical sequences of purchased baskets to recommend next basket that satisfies user's different intentions.

In another line of work, a K-nearest neighbor (KNN) based approach (Hu et al.) [1] was proposed to exploit personalized item frequency information for improving the performance of NBRs. Similarly, another approach based on KNN and collaborative filtering (CF) (Faggioli et al.) [19] was proposed

✉ Wenpeng Lu is the corresponding author.

to model the recency of items for NBRs.

Although there are many studies on NBR existing in the literature and most of them have achieved great success, one critical issue has attracted much attention: there is a lack of a systematic and unified framework to comprehensively and systematically categorize NBR approaches and evaluate them in a rigorous way. As a result, it is not very clear what is the latest research progress in this vibrant area and how the different NBR approaches really perform. There is also a lack of problem formalization and unified experimental settings for NBR research, making it hard to compare different approaches in a fair way [20]. Also, the inconsistency on the selection of datasets, baselines commonly exist in various studies.

The aforementioned gaps in the existing NBR research motivate us to conduct a systematic investigation on various NBR studies in the literature in both a quality and a quantity way. To be specific, in this work, we first provide a formal problem statement for the NBR problem, then summarize the research progress in NBR area by categorizing and comparing the representative and state-of-the-art NBR approaches. Afterwards, we set up a unified experimental environment and conduct a systematic empirical study on most representative NBR approaches which are carefully selected.

To the best of our knowledge, this is the first work to systematically investigate the NBR approaches in the literature. The main contributions of this work are summarized below:

- We provide a novel taxonomy to well categorize and organize a variety of representative NBR approaches. As a result, an overview of the current progress of research in NBR area has been provided.
- We conduct a comprehensive and systematic empirical study on the representative and state-of-the-art NBR approaches. This provide a unified evaluation for various NBR approaches to well compare their performance in a rigorous and fair way. We hope this work shade some light for the future studies in this vibrant area.

II. RELATED WORK

In this section, we first clarify the difference between next-basket recommendation and bundle recommendation which is a task quite relevant to but different from next-basket recommendation. Then, we review some existing surveys and reviews related to the topic of next-basket recommendation.

A. Next Basket Recommendation vs. Bundle Recommendation

There have been some studies on other tasks relevant to next-basket recommendation, among which, bundle recommendation [21] is the most typical one. Although superficially similar, next-basket recommendation and bundle recommendation vary in settings and assumptions while researchers often mix up them. Therefore, it is necessary to conduct a comparison between them and thus differentiate next-basket recommendation from bundle recommendation.

Next basket recommendation and bundle recommendation are built on basket data and bundle data respectively. So it is necessary to first clarify the difference between basket data and bundle data. A basket is a list of items purchased by a user in one shopping visit. There is usually no clear order over the items within a basket. The items in one basket may be correlated or uncorrelated, which depends on the specific cases. Given a sequence of baskets, they are often sorted in ascending order in terms of a shopping visit. In comparison, a bundle is the integration of two or more highly correlated items [22]. The items in a bundle are unordered and are often similar or complementary with each other [23]. When regarding bundles as a sequence, bundle recommendation usually gives the order by sorting the bundles in terms of their prices in a descendant order.

With the user's transaction history, namely a sequence of historical baskets, next basket recommendation is often formalized as a sequential prediction task, i.e., next basket recommendation targets to predict what items the user will buy in her/his next shopping visit. In such a case, the sequential dependence over a user's historical purchased baskets was fully modelled and employed. For example, when we know a user has purchased a printer in her current shopping basket, we can recommend some relevant items such as printing paper and toners as her next basket. In comparison, bundle recommendation [24] is formalized as an optimization problem which aims to select a set of optimal and correlated items from the massive candidate items to accomplish a certain consumption goal of a user. Different from next-basket recommendation, no obvious sequential dependency will be considered and modelled in bundle recommendation. For instance, given the high relevance between iPhone and AirPods, we can recommend them together as a bundle to a user who likes Apple products without modeling any sequential purchase behaviors.

B. Related Surveys

There are also many other studies related to next-basket recommendation, e.g., session-based recommendation and sequential recommendation. For session based recommendation, Wang et al. [5], [25] provided a comprehensive and systematic survey to formally define the research problems, illustrate the main research challenges, review the main research progress and point out the promising research directions in this area. Ludewig et al. [26] provided a systematic empirical study on a variety of session-based recommendation algorithms to comprehensively compare their recommendation performance. For sequential recommendation, Wang et al. [27] summarized the key challenges, progress and future directions in this important research area. Fang et al. [28] summarized and compared deep learning based approaches for sequential recommendation.

However, although relevant, session-based recommendation and sequential recommendation are totally different from next-basket recommendation. The reason is that session-based recommendation and sequential recommendation are based on session data and sequence data respectively, in which a set of items (a session) or a sequence of items are taken as the input of the recommendation algorithm. As a result, there is no clear basket structure inside session data or sequence data. In another words, they mainly recommend the next item within

the same basket by modeling the intra-basket dependencies. In contrast, the input data for next-basket recommendation is a sequence of baskets, and an NBR aims to recommend the next basket via modeling the inter-basket dependencies.

Although a series of review work and empirical studies have been done in the areas of session-based recommendation and sequential recommendation [29]–[31]. There is no systematic work to provide a comprehensive investigation on the various studies in the NBR area and there is not a unified and rigorous evaluation on the different NBR algorithms. Given the increasing popularity and potential of NBR and the emerging research progress in this area, a systematic and thorough summarization, evaluation and analysis of NBR approaches is in urgent demand. As the first attempt, this paper explores the field of NBR with an emphasis on the problem statement, classes of existing NBR approaches, benchmarking evaluation and provides further insights of future prospects in the area.

III. PROBLEM STATEMENT

In this section, we formally define the research problem of NBR and then discuss the main work mechanism of NBR.

There are different definitions on NBR task based on different domains in the collected papers. For example, a basket could be defined not only as a series of products purchased in one transaction event in E-commerce domain, but also as a set of places visited in a trip [15], [32] in the tourism domain. To simplify the concepts, we call both a product and a place as an item in this paper. Consequently, a *basket* is defined as a collection of items which have been purchased together in one transaction event by a given user. In the next-basket recommendation task, often given a sequence of historical baskets purchased by a user recently as the input, a next-basket recommender system is built and trained on such input data to predict the next basket of items which mostly interest the user via modeling the inter-basket dependencies. Usually, the prediction is performed in the form of generating a personalized ranked list of items (as depicted in Figure 1), which are supposed to form the next basket for the given user.

Now we follow the problem statement in [17] to formally define the research problem of next-basket recommendation. Given a transaction dataset $D = \{s_1, ..., s_{|D|}\}$ where $|D|$ is the total number of sequences, it contains a set of sequences of shopping baskets (called baskets for short). In D, each sequence $s = \{b_1, ..., b_{|s|}\}(s \in D)$ consists of a list of historical baskets purchased by a certain user and they are sorted in the order of purchase time. In each sequence, a basket $b = \{v_1, ..., v_{|b|}\}(b \in s)$ contains a collection of items which were purchased in one transaction event. All the items occurring in the whole dataset constitute the universal item set $V = \{v_1, ..., v_{|V|}\}$ while all the users occurring in the dataset form the universal user set $U = \{u_1, ..., u_{|U|}\}$.

Given a sequence of baskets $s = \{b_1, ..., b_t\}$, we pick up one basket, usually the last one b_t as the target basket to be predicted, while all the other baskets occurring prior to b_t will be taken as the corresponding context, denoted as $C_t = \{b_1, ..., b_{t-1}\}$. Note that in the prediction and recommendation stage, the target basket is unknown and needs to be predicted by the next-basket recommender systems. In a next-basket recommendation task, given a context C_t of a user u, namely the purchase history of u, a next-basket recommender system aims to predict the corresponding user's choices for the t^{th} basket, namely to generate a list of items which are most probably to appear in the user's next basket b_t. This can be formally defined below:

$$b_t = f(C_t, u). \tag{1}$$

IV. CLASSES AND COMPARISON OF NBR APPROACHES

In this section, to provide an overview of the NBR research progress, we first propose a taxonomy to well classify the representative and state-of-the-art NBR approaches, and then compare the different classes of approaches systematically. Specifically, there are a variety of studies in NBR area and it is impossible to analyze each of them in detail. Therefore, in this section, we carefully select 13 most representative and highly cited works to analyze. To be specific, we first search the papers whose title contain the keyword *next basket recommendation* in Google Scholar and then select those papers with equal or more than 10 citations. To provide a more straightforward view on the research progress in NBR, we conduct a bibliometric analysis over the selected literature. The distributions of publication year, experimental datsets, compared approaches and used evaluation metrics of the literature have been shown in Figure 2.

A. A Categorization of NBR Approaches.

As depicted in Figure 3, according to the utilized data mining or machine learning models and techniques, three major classes for NBR approaches are identified from the literature, i.e., (1) conventional NBR approaches which are built on conventional data mining approaches such as pattern mining and K-nearest neighbour (KNN); (2) latent representation approaches which are mainly built on representation learning techniques such as latent embedding; and (3) deep neural network approaches which are typically built on deep learning models including RNN. These three classes can be further divided into eight sub-classes. Specifically, conventional NBR approaches contain four sub-classes: pattern mining approaches, K-nearest neighbour approaches, and Markov chain based approaches; latent representation method mainly contains one sub-class, i.e., distributed representation; and

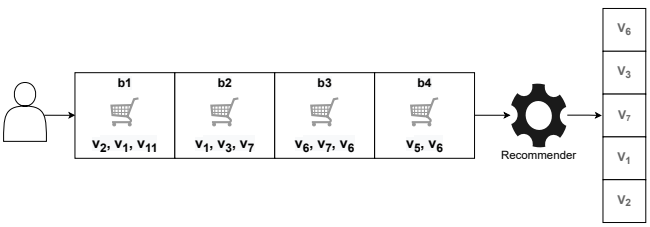
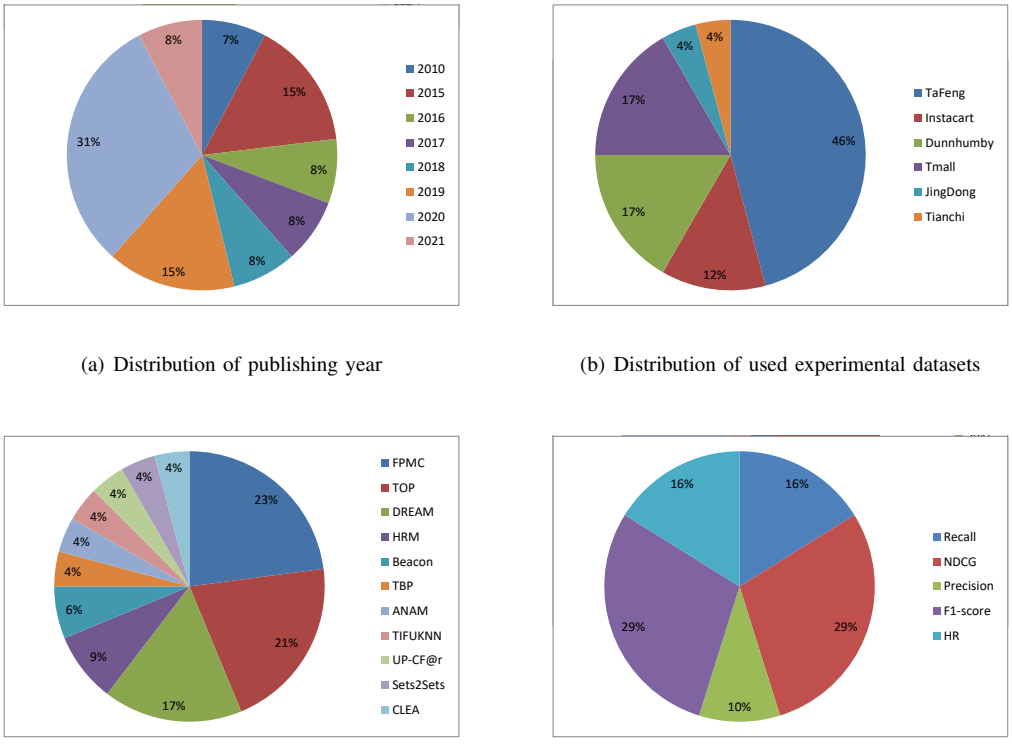


Fig. 1: A running example for next basket recommendation.



(a) Distribution of publishing year

(b) Distribution of used experimental datasets

(c) Distribution of compared approaches

(d) Distribution of used evaluation metrics

Fig. 2: (a) shows the distribution of publishing year of collected papers ; (b) depicts the popularity of top-6 datasets; (c) shows the popularity of 11 compared NBR approaches; (d) shows the popularity of top-5 evaluation metrics.

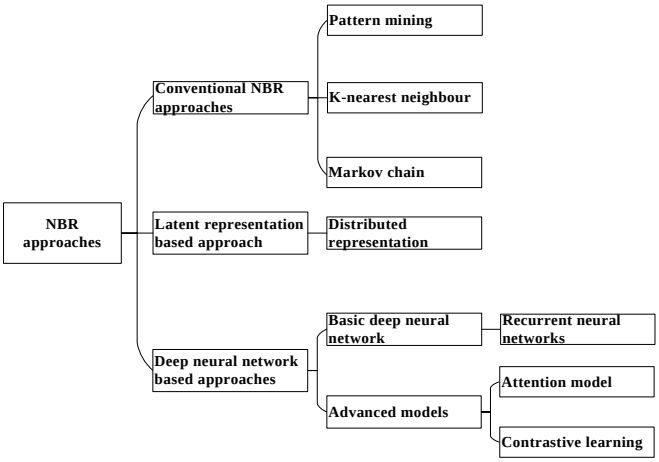


Fig. 3: The classes of NBR approaches.

deep neural network approaches contain two sub-classes: basic deep neural networks and advanced models. In addition, the sub-class of basic deep neural network mainly contains recurrent neural network based NBR. The sub-class of advanced models contains two sub-subclass, i.e., attention models and contrastive learning. In particular, except the approaches based on a single algorithm/model, there are some hybrid approaches which combine more than one algorithms/models. For example, ANAM [16] combines recurrent neural network and attention model and CLEA [2] combines a contrastive learning framework and a recurrent neural network, etc.

B. A Comparison of Different Classes of NBR Approaches

Generally, conventional NBR approaches are straightforward and not so complicated since they are mostly based on conventional models and algorithms like pattern mining [33], KNN models [1] and Markov chain models [10]. As a result, they are mostly more efficient and easy to implement. And they are more likely to be applicable to simple datasets in which the dependencies within or between baskets are not so complex. Furthermore, although simple, they are sometimes quite effective for NBR tasks. For exampled, the TIFUKNN [1], a K-nearest neighbour (KNN) based approach, can even outperform some deep learning based approaches like Dream and Beacon in terms of most of the metrics on the three experimental datasets (cf. TABLE II).

In contrast, deep neural network based approaches are relatively sophisticated, which utilize various neural architectures to capture the intrinsic features and dependencies within and between baskets. As a result, they are usually more time consuming. In some cases, they may not perform as stable as the conventional approaches do. In most cases, deep learning based approaches can better capture the complex intra- and inter-basket dependencies and thus can achieve better recommendation performance. For example, recurrent

neural networks (RNN) have been verified as an effective solution in handling sequences of baskets in recent years [13]–[16].

Last but not least, latent representation based approaches usually do not employ complex deep neural architecture, and thus they are often more efficient and simpler compared with deep learning based approaches. On the other hand, compared with conventional approaches, due to the power of utilized embedding techniques, latent representation based approaches are easier to capture some implicit and hidden patterns in the data and thus are likely to achiever better performance. A typical example in this class is HRM [12], which is a distributed representation based NBR approach. It employs a three-layer structure to construct a representation of a user's last basket to predict his/her next basket.

In order to have a better understanding of the representative NBR approaches from different classes and how each category of approaches contribute to promote the evolution of NBR, we systematically compare the selected representative NBR approaches from various classes. We analyze the pros and cons of each approach, which are described in Table I.

V. Datasets and Compared Approaches

A. Datasets

Through the analysis of datasets in the collected papers, we find two major issues: (1) Due to user privacy limitation, some datasets are not publicly available; (2) For some datasets, though they share the same name, they have different versions actually. For example, we find more than two versions for Jing-Dong dataset, which is provided by the recommender system related competition held by JingDong Retail Group, and is updated each year. Overall, as the purposes and requirements of collected papers are different, different datasets are adopted. Figure 2(b) shows the popularity of top-6 datasets, where non-public datasets are not counted. Considering the popularity and characteristics of the datasets above, we select 3 datasets in our experiments, which are commonly used to evaluate the performance of next-basket prediction [17], described as follows:

- **TaFeng**[1] released on Kaggle, which contains 4 months of shopping transactions with 32,266 users and 23,812 items. This dataset is a Chinese grocery store dataset with numerous baskets of purchased items. All data in the dataset is utilized in our experiments.
- **Instacart**[2] released on Instacart challenge, which contains over 3 million online purchases from more than 200,000 users. Following the work of Qin et al. [2], we conduct our experiments by randomly sampling 10% of the user transaction records from the test user set.
- **Dunnhumby**[3] released by Dunnhumby, a business data processing and analysis company. It records more than 2 years of purchases of 2,500 households who are frequent shoppers at a retailer. Following the work of Yu et al. [34], we adopt the transactions in the first two months to conduct our experiments.

The above datasets are publicly available with the information about which user purchases which item at which time, which are suitable to be adopted as benchmark datasets on the task of NBR.

B. Data Pre-processing

As the original datasets are always huge and sparse, data pre-processing is necessary for the following evaluations. Generally, the infrequent items and inactive users with fewer interactions will be removed during data-processing procedures. By analyzing the collected papers, we find that all papers adopt pre-processing strategies while the strategies they adopt are not uniform. To better fairly evaluate the performance of different approaches across different data, we pre-process the datasets and filter out the items that were purchased less than n times. The values of n are set to 10, 20 and 17 for TaFeng, Instacart, and Dunnhumby datasets, respectively. For different datasets, we construct baskets according to its data characteristics. In TaFeng, it contains the timestamp of shopping transactions from 2020-11-01 to 2020-02-28, so we adopt one day as time unit, i.e., the items in the same day count as a basket. In Instacart, we treat all items purchased in the same order as a basket. Similarly, items purchased in the same transaction are treated as a basket in Dunnhumby. For each user, we arrange his/her baskets in chronological order to form a sequence. We only reserve the baskets with the size between 2 to 60 in each dataset. The statistic information of all the datasets after pre-processing is shown in Table II.

Following the work of Hu et al. [1], we partition the data into 5 folds across sequences, 4 folds is utilized for training, and 1 fold is utilized for testing. We further reserve the data of 10% sequences in the training set as the validation set to tune hyper-parameters.

C. Compared Approaches

In our collected papers, the compared approaches adopted by each paper are not same. We summarize the top-11 widely-compared approaches, as shown in Figure 2(c). Furthermore, considering their reproducibility and popularity, we select 8 approaches in our experiments. Among them, Sets2Sets is an approach for sequential set prediction with an encoder-decoder framework, which is modified to be applicable for NBR. The details of the eight approaches are described as follows:

- **TOP**: a non-personalized approach which recommends the most popular items according to their purchase frequencies [35].
- **FPMC**: a Markov Chain (MC) based approach which combines MC with Matrix Factorization (MF) to model the pairwise item-item transition patterns from adjacent baskets to recommend the next basket of items [10].
- **DREAM**: an recurrent neural network (RNN) based approach to consider both user's dynamic interest and

[1]https://www.kaggle.com/chiranjivdas09/ta-feng-grocery-dataset
[2]https://www.kaggle.com/c/instacart-market-basket-analysis/data
[3]https://www.dunnhumby.com/source-files/

TABLE I: A comparison of different classes of NBR approaches.

Approach	Class	Pros	Cons
TBP [33]	Pattern mining	TBP captures different factors influencing user's next choices.	TBP is usually biased to frequent items while ignoring less-frequent ones.
TIFUKNN [1]	KNN	TIFUKNN exploits personalized item frequency information.	They only recommend items that a user has purchased in the past.
UP-CF@r [19]	KNN	UP-CF@r considers the recency of items in the purchased history.	
FPMC [10]	MC	FPMC models user's long-term preferences and the transition patterns of items.	FPMC only captures the first-order dependencies while ignoring the higher-order ones.
NN-Rec [11] HRM [12]	Representation	They predict next basket based on the representations of users and baskets.	They fail to capture higher-order dependencies among baskets.
DREAM [13]	RNN	DREAM learns a dynamic interests of a user and captures global sequential features of all baskets.	They may generate false dependencies and are hard to capture item frequency features.
ANAM [16]		ANAM utilizes the attention mechanism to integrate items and their category information.	
Beacon [15]		Beacon utilizes correlation information over items.	
Sets2Sets [14]		Sets2Sets employs an encoder-decoder architecture with repeated purchase pattern.	
IntNet [17] Int2Ba [18]		They consider the human's intentions contained in the purchased history.	
CLEA [2]		CLEA denoises baskets and extracts credibly relevant items to enhance NBR.	

TABLE II: The statistics of experimental datasets.

Statistics	TaFeng	Instacart	Dunnhumby
#Users	20,212	19,982	22,530
#Baskets	105,140	280,941	214,861
#Items	10,411	13,400	3,920
#Basket/user	5.20	14.06	9.53
#Items/basket	6.32	9.61	7.45

The rows #Users, #Baskets, #Items, #Basket/user, #Items/basket correspond to the number of users, the number of baskets over all users, the number of items, the average number of baskets per user and the average number of items per basket, respectively.

global sequential features that reflect interactions among baskets [13].

- **Beacon**: an RNN based approach which encodes the basket sequence to model the pairwise correlations among items [15].
- **Sets2Sets**: an RNN based encoder-decoder approach for set/basket prediction. In addition, an attention mechanism focusing on item frequency is proposed to improve the performance [14].
- **UP-CF@r**: a simple approach that relies on the user-wise popularity with collaborative filtering and the recency of shopping [19].
- **TIFUKNN**: a K-nearest neighbour (KNN) based approach that exploits personalized frequency information of items. A novel repeated purchase pattern is applied in this approach [1].
- **CLEA**: an RNN based contrastive learning approach to denoise basket generation by identifying relevant items in the history [2].

VI. Benchmarking Evaluation

A. Evaluation Metrics

In our collected papers, the evaluation metrics adopted by each paper change greatly. Figure 2(d) shows the popularity of evaluation metrics. In order to comprehensively and fairly evaluate different approaches, we adopt the following 7 metrics: Recall, Precision, F1-Score, Person-wise Hit Ratio (PHR), Normalized Discounted Cumulative Gain (NDCG), Mean Average Precision (MAP) and Mean Reciprocal Rank (MRR), where the latter two metrics are not reflected in the collected papers, but they are common in the top-K recommendation. The details of the seven metrics are introduced as follows:

- **Recall**: it is a widely-used metric in NBR [36], which measures the proportion of ground-truth items in a predicted basket that are correctly recommended. Recall is calculated by:

$$\text{Recall@K} = \frac{|S' \cap S|}{|S|},$$

where S and S' are the ground-truth items and the top-K items of predicted basket respectively, $|S|$ denotes the size of ground-truth items.

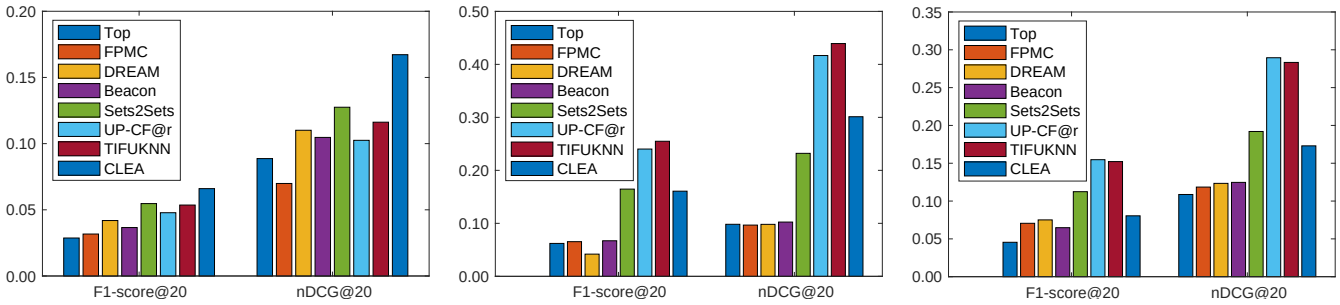
- **Precision**: it is corresponding with Recall, measures the proportion of ground-truth items in the top-K items of predicted baskets [36], which is calculated by:

$$\text{Precision@K} = \frac{|S' \cap S|}{K}.$$

- **F1-Score**: it is the harmonic mean of precision and recall

TABLE III: The statistics of compared approaches.

Dataset				TaFeng				
K	Approaches	Recall	Precision	F1-score	PHR	MAP	MRR	NDCG
5	TOP	0.0670	0.0420	0.0412	0.1870	0.1426	0.1438	0.0733
	FPMC	0.0632	0.0433	0.0514	0.1912	0.1066	0.1070	0.0557
	DREAM	0.0921	0.0533	0.0675	0.2329	0.1580	0.1608	0.0891
	Beacon	0.0820	0.0489	0.0493	0.2117	0.1473	0.1510	0.0817
	Sets2Sets	0.0951	0.0690	0.0647	0.2860	0.1718	0.1763	0.0925
	UP-CF@r	0.0741	0.0644	0.0565	0.2601	0.1539	0.1574	0.0726
	TIFUKNN	0.0833	0.0713	0.0624	0.2799	0.1626	0.1663	0.0790
	CLEA	**0.1280**	**0.0900**	**0.0842**	**0.3547**	**0.2321**	**0.2386**	**0.1284**
10	TOP	0.0768	0.0269	0.0326	0.2282	0.1455	0.1495	0.0786
	FPMC	0.0755	0.0274	0.0402	0.2314	0.1096	0.1122	0.0616
	DREAM	0.1170	0.0362	0.0553	0.3060	0.1648	0.1705	0.1004
	Beacon	0.1059	0.0348	0.0435	0.2863	0.1543	0.1608	0.0930
	Sets2Sets	0.1315	0.0512	0.0614	0.3797	0.1730	0.1816	0.1082
	UP-CF@r	0.1058	0.0496	0.0560	0.3483	0.1600	0.1691	0.0889
	TIFUKNN	0.1236	0.0537	0.0619	0.3765	0.1680	0.1791	0.0982
	CLEA	**0.1609**	**0.0648**	**0.0760**	**0.4446**	**0.2353**	**0.2514**	**0.1472**
Dataset				Instacart				
K	Approaches	Recall	Precision	F1-score	PHR	MAP	MRR	NDCG
5	TOP	0.0487	0.0956	0.0581	0.3668	0.2270	0.2321	0.0666
	FPMC	0.0481	0.0948	0.0620	0.3600	0.2211	0.2264	0.0651
	DREAM	0.0763	0.0486	0.0594	0.2176	0.1543	0.1481	0.0754
	Beacon	0.0539	0.1049	0.0638	0.3930	0.2181	0.2250	0.0695
	Sets2Sets	0.1266	0.1652	0.1209	0.5503	0.3141	0.3260	0.1375
	UP-CF@r	0.2512	0.3580	0.2512	0.7946	0.5818	0.6138	0.2970
	TIFUKNN	**0.2616**	**0.3733**	**0.2619**	**0.8052**	**0.5992**	**0.6312**	**0.3094**
	CLEA	0.1850	0.3025	0.1973	0.7136	0.5623	0.5865	0.2450
10	TOP	0.0731	0.0735	0.0655	0.4614	0.2223	0.2448	0.0832
	FPMC	0.0712	0.0716	0.0697	0.4508	0.2182	0.2388	0.0809
	DREAM	0.1012	0.0359	0.0530	0.3037	0.1543	0.1596	0.0874
	Beacon	0.0767	0.0769	0.0686	0.4755	0.2174	0.2362	0.0851
	Sets2Sets	0.2058	0.1463	0.1710	0.7157	0.3155	0.3550	0.1862
	UP-CF@r	0.3480	0.2731	0.2650	0.8590	0.5448	0.6226	0.3613
	TIFUKNN	**0.3698**	**0.2896**	**0.2812**	**0.8756**	**0.5618**	**0.6409**	**0.3805**
	CLEA	0.2135	0.1857	0.1738	0.7439	0.5020	0.5659	0.2537
Dataset				Dunnhumby				
K	Approaches	Recall	Precision	F1-score	PHR	MAP	MRR	NDCG
5	TOP	0.0966	0.0763	0.0606	0.3294	0.2190	0.2282	0.0861
	FPMC	0.0746	0.1033	0.0866	0.3928	0.2566	0.2674	0.0903
	DREAM	0.0756	0.1049	0.0879	0.3984	0.2616	0.2726	0.0918
	Beacon	0.0795	0.1038	0.0737	0.3948	0.2517	0.2682	0.0932
	Sets2Sets	0.1040	0.1249	0.0904	0.4353	0.2515	0.2867	0.1112
	UP-CF@r	**0.1794**	**0.2417**	**0.1693**	**0.6005**	**0.4367**	**0.4602**	**0.2105**
	TIFUKNN	0.1725	0.2329	0.1630	0.5947	0.4260	0.4465	0.2027
	CLEA	0.1193	0.1720	0.1165	0.5042	0.3512	0.3663	0.1422
10	TOP	0.1169	0.0566	0.0555	0.4161	0.2194	0.2385	0.0983
	FPMC	0.0961	0.0705	0.0813	0.4560	0.2483	0.2758	0.1037
	DREAM	0.1006	0.0732	0.0847	0.4711	0.2535	0.2822	0.1069
	Beacon	0.1030	0.0726	0.0709	0.4667	0.2493	0.2776	0.1079
	Sets2Sets	0.1695	0.1102	0.1091	0.5751	0.2563	0.2867	0.1488
	UP-CF@r	**0.2480**	**0.1795**	**0.1733**	**0.6764**	**0.4149**	**0.4703**	**0.2525**
	TIFUKNN	0.2419	0.1734	0.1677	0.6713	0.4054	0.4568	0.2444
	CLEA	0.1476	0.1102	0.1048	0.5555	0.3555	0.3861	0.1673



(a) TaFeng (b) Instacart (c) Dunnhumby

Fig. 4: The performance of different approaches in terms of F1-Score and NDCG.

[12], which is calculated by:

$$\text{F1} - \text{Score@K} = \frac{2 \times \text{Precision@K} \times \text{Recall@K}}{\text{Precision@K} + \text{Recall@K}}.$$

- **PHR**: it is person-wise hit ratio which represents the ratio of users whose ground-truth items appear in the recommendation list [34]. PHR is calculated by

$$\text{PHR@K} = \frac{1}{N}\sum_{i=1}^{N} hr(i),$$

where

$$hr(i) = \begin{cases} 1, & |S' \cap S| > 0 \\ 0, & |S' \cap S| = 0, \end{cases}$$

and N denotes the number of testing users.

- **NDCG**: it is a ranking based measure, which focuses on the order of items in a predicted basket [34]. NDCG is more sensitive to higher ranked items. That is to say, higher NDCG indicates the ground-truth items are recommended at higher ranks. NDCG is defined as:

$$\text{NDCG@K} = \frac{1}{\sum_{j=1}^{|S|} \frac{1}{\log_2(j+1)}} \sum_{i=1}^{K} \frac{\partial(S', S)}{\log_2(k+1)},$$

where $\partial(S', S)$ returns 1 when the item in the predicted basket appears in the ground-truth, otherwise 0.

- **MAP**: it is a relatively common evaluation metric [35]. Average Precision (AP) is calculated in the following way:

$$\text{AP}(i) = \frac{1}{m}\sum_{i=1}^{m} m \times \frac{1}{p_i},$$

where m is the number of ground-truth items if they appear in the recommendation list ($m <= K$), and p_i is the position of item i in the recommendation list. And, MAP is the mean of AP, which is calculated by:

$$\text{MAP@K} = \frac{1}{N}\sum_{i=1}^{N} AP(i),$$

where N is same to that in PHR.

- **MRR**: it is derived from the information retrieval. Reciprocal Rank(RR) refers to the inverse of the ranking of correctly recommended item in the top-K recommendation list [35]. MRR is the mean of Reciprocal Rank, which is calculated by:

$$\text{MRR@K} = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{q_i},$$

where we define q_i as the ranking of the first item that appears in the ground-truth in the top-K recommendation list.

Following the settings of traditional NBR work [1], [2], for each approach, we will recommend a predicted basket (i.e., recommendation list) with a fixed size K for evaluation. All the metrics are calculated across all predicted baskets, and all metrics have the same characteristic: the larger value, the better performance.

B. Performance of Compared Approaches

In order to provide a better reference for fair comparison, Table III shows the performance of eight approaches across seven metrics on the three datasets [4].

Finding 1: *Although conventional approaches are simple, they can achieve wonderful performance.* Among the conventional approaches including Top, FPMC, UP-CF@r and TIFUKNN, the two former ones perform similarly on all datasets. As the simplest approach, TOP even performs better than FPMC w.r.t Recall, MAP, MRR and NDCG in most cases. This result confirms the importance of item frequency information in next-basket recommendation. Besides, this demonstrates that TOP should always serve as a baseline when a new model is proposed [38]. On Instacart and Dunnhumby, UP-CF@r and TIFUKNN are the best performing approaches on all metrics. The two approaches are similar in that they both combine the attributes of items (i.e., recency and frequency) with a user-based nearest neighbor idea. This implies that though the conventional approaches are simple, they are also able to show powerful ability on suitable datasets.

Finding 2: *The deep neural network based approaches fail to demonstrate consistent and stable performance.* Among the

[4]The approaches in the table are sorted by their publication year.

TABLE IV: A list of representative open-source NBRs algorithms

Algorithm	Utilized model	Venue	Link
DREAM [13]	RNN	SIGIR 2016	https://github.com/yihong-chen/DREAM
Sets2Sets [14]	RNN	KDD 2019	https://github.com/HaojiHu/Sets2Sets
Beacon [15]	RNN	IJCAI 2019	https://github.com/PreferredAI/beacon
CLEA [2]	RNN	SIGIR 2021	https://github.com/QYQ-bot/CLEA
MBN [37]	RNN	TKDD 2022	https://github.com/gybuay/MBN
HRM [12]	Distributed representation	SIGIR 2015	https://github.com/chenghu17/Sequential_Recommendation
ReCANet [38]	Distributed representation	SIGIR 2022	https://github.com/mzhariann/recanet
UP-CF@r [19]	KNN	UMAP 2020	https://github.com/MayloIFERR/RACF
TIFUKNN [1]	KNN	SIGIR 2020	https://github.com/HaojiHu/TIFUKNN
DNNTSP [34]	GNN	KDD 2020	https://github.com/yule-BUAA/DNNTSP
MITGNN [39]	GNN	BigData 2020	https://github.com/JimLiu96/MITGNN
FPMC [10]	MC	WWW 2010	https://github.com/khesui/FPMC
TBP [33]	Pattern mining	ICDM 2017	https://github.com/GiulioRossetti/tbp-next-basket

TABLE V: A list of commonly used and publicly accessible real-world datasets for NBRs

Dataset	#Users	#Items	#Baskets	Avg. basket size	#Baskets per user	Reference
TaFeng	12,805	10,829	89,543	6.39	6.99	[12] [13] [33] [16] [14] [15] [2] [1] [34]
Instacart	7,282	12,515	115,717	9.63	15.89	[19] [7] [39] [2] [38]
Dunnhumby	2,488	26,779	269,951	9.02	108.50	[14] [38] [2] [1] [19]
Tmall[5]	102,681	36,113	739,178	1.72	7.19	[12] [13] [18] [17] [37]
JingDong[6]	60,534	41,186	243,769	1.16	4.02	[16] [37]
Tianchi[7]	6,924	27,637	34,749	2.12	5.01	[37]
Valuedshopper[8]	9,997	6,421	280,762	9.17	28.08	[2] [38] [1]
Taobao[9]	47,392	90,440	169,840	1.48	3.58	[34]

[5] https://tianchi.aliyun.com/dataset/dataDetail?dataId=42
[6] https://jdata.jd.com/html/detail.html?id=8
[7] https://tianchi.aliyun.com/competition/entrance/231522/information
[8] https://www.kaggle.com/c/acquire-valued-shoppers-challenge/overview
[9] https://tianchi.aliyun.com/dataset/dataDetail?dataId=649

deep neural network based approaches including DREAM, Beacon, Sets2Sets and CLEA, CLEA can automatically capture interactions between historical items and the target item, which achieves the best performance on TaFeng, but it is not the best one for any of the other two datasets. We suspect that it is related to the characteristics of datasets because the average basket size (i.e., items/basket) of TaFeng is lower than those of Instacart and Dunnhumby (see from Table II). Sets2Sets utilizes repeated pattern based on encoder-decoder architecture, which outperforms DREAM, Beacon and even CLEA in Dunnhumby. This indicates that there exist numerous repeated purchases in NBR task.

Finding 3: *The conventional approaches can beat the deep neural network based ones on most datasets.* Regarding the different compared approaches, we can observe that their performances change a lot across datasets, and there is no approach outperforms all others on each dataset. So we further analyze the performance on three datasets in terms of two key metrics, i.e., F1-Score and NDCG, as shown in Figure 4. We find that Sets2Sets, UP-CF@r, TIFUKNN and CLEA stand out among other approaches. Though Sets2Sets and CLEA show better performance than UP-CF@r and TIFUKNN on Tafeng dataset, they are beaten on Instacart and Dunnhumby datasets. Sets2Sets and CLEA are deep neural based approaches, while UP-CF@r and TIFUKNN are conventional approaches based on KNN. Considering the basket length of Tafeng is shorter than Instacart and Dunnhumby, this may indicate that Sets2Sets and CLEA are more applicable on the datsets with shorter baskets, however, UP-CF@r and TIFUKNN can effectively recommend next basket on the datasets with longer baskets.

According to the experimental data, we find that under same experimental settings and on the same datasets, the performances of deep neural approaches are inferior to those of conventional ones in most cases.

VII. NBR Algorithms and Datasets

For facilitating the access for empirical analysis, in table IV, we summarize source codes of algorithms for NBRs which utilize different models. The listed NBR algorithms are publicly accessible and commonly used as baselines in existing work. In addition to algorithms, datasets are necessary for evaluating NBRs algorithms. In order to facilitate further analysis of the investigated algorithms, in table V, we also summarize datasets for NBRs which can be built as baskets in ascending order in terms of users' shopping visits.

VIII. Conclusion

In this paper, we have systematically investigated and evaluated the representative works on NBR task. Specifically, we clarify the difference between NBRs and traditional RSs, and formulate the problem statement of NBR. Then, we categorize the existing work into three groups, i.e., conventional approaches, latent representation based approaches and deep neural network based approaches, whose advantages and disadvantages are analyzed. With the aim of better understanding and systematical evaluation on NBRs, we further analyze the popular NBR algorithms and evaluation metrics, and re-run them on the same dataset with unified experimental settings to compare their performance. We have provided a unified framework to fairly evaluate different NBR approaches, which can be utilized as a valuable reference for the related research on NBR. We hope that this work can make readers understand NBR more clearly and make the evaluation of NBRs more fairly and effectively.

Acknowledgment

The work is partly supported by National Nature Science Foundation of China (61502259), and Key Program of Science and Technology of Shandong (2020CXGC010901), and Studio Project of the Research Leader in Jinan (2019GXRC062).

References

[1] H. Hu, X. He, J. Gao, and Z. Zhang, "Modeling personalized item frequency information for next-basket recommendation," in *SIGIR*, 2020, pp. 1071–1080.

[2] Y. Qin, P. Wang, and C. Li, "The world is binary: Contrastive learning for denoising next basket recommendation," in *SIGIR*, 2021, pp. 859–868.

[3] W. Lu, R. Wang, S. Wang, X. Peng, H. Wu, and Q. Zhang, "Aspect-driven user preference and news representation learning for news recommendation," *IEEE Transactions on Intelligent Transportation Systems*, 2022.

[4] Q. Zhang, S. Wang, W. Lu, C. Feng, X. Peng, and Q. Wang, "Rethinking adjacent dependency in session-based recommendations," in *PAKDD*, 2022, pp. 301–313.

[5] S. Wang, L. Cao, Y. Wang, Q. Z. Sheng, M. A. Orgun, and D. Lian, "A survey on session-based recommender systems," *ACM Computing Surveys*, vol. 54, no. 7, pp. 1–38, 2021.

[6] W. Song, S. Wang, Y. Wang, and S. Wang, "Next-item recommendations in short sessions," in *RecSys*, 2021, pp. 282–291.

[7] N. Wang, S. Wang, Y. Wang, Q. Z. Sheng, and M. Orgun, "Modelling local and global dependencies for next-item recommendations," in *WISE*, 2020, pp. 285–300.

[8] S. Wang, L. Hu, Y. Wang, Q. Z. Sheng, M. Orgun, and L. Cao, "Modeling multi-purpose sessions for next-item recommendations via mixture-channel purpose routing networks," in *IJCAI*, 2019, pp. 3771–3777.

[9] S. Wang, L. Hu, L. Cao, X. Huang, D. Lian, and W. Liu, "Attention-based transactional context embedding for next-item recommendation," in *AAAI*, 2018, pp. 2532–2539.

[10] S. Rendle, C. Freudenthaler, and L. Schmidt-Thieme, "Factorizing personalized markov chains for next-basket recommendation," in *WWW*, 2010, pp. 811–820.

[11] S. Wan, Y. Lan, P. Wang, J. Guo, J. Xu, and X. Cheng, "Next basket recommendation with neural networks," in *RecSys (Poster)*, 2015.

[12] P. Wang, J. Guo, Y. Lan, J. Xu, S. Wan, and X. Cheng, "Learning hierarchical representation model for next-basket recommendation," in *SIGIR*, 2015, pp. 403–412.

[13] F. Yu, Q. Liu, S. Wu, L. Wang, and T. Tan, "A dynamic recurrent model for next basket recommendation," in *SIGIR*, 2016, pp. 729–732.

[14] H. Hu and X. He, "Sets2Sets: Learning from sequential sets with neural networks," in *SIGKDD*, 2019, pp. 1491–1499.

[15] D.-T. Le, H. W. Lauw, and Y. Fang, "Correlation-sensitive next-basket recommendation," in *IJCAI*, 2019, pp. 2808–2814.

[16] T. Bai, J.-Y. Nie, W. X. Zhao, Y. Zhu, P. Du, and J.-R. Wen, "An attribute-aware neural attentive model for next basket recommendation," in *SIGIR*, 2018, pp. 1201–1204.

[17] S. Wang, L. Hu, Y. Wang, Q. Z. Sheng, M. Orgun, and L. Cao, "Intention Nets: Psychology-inspired user choice behavior modeling for next-basket prediction," in *AAAI*, 2020, pp. 6259–6266.

[18] S. Wang, L. Hu, Y. Wang, Q. Z. Sheng, M. Orgun, and et al., "Intention2Basket: A neural intention-driven approach for dynamic next-basket planning," in *IJCAI*, 2020, pp. 2333–2339.

[19] G. Faggioli, M. Polato, and F. Aiolli, "Recency aware collaborative filtering for next basket recommendation," in *UMAP*, 2020, pp. 80–87.

[20] S. Wang, X. Zhang, Y. Wang, H. Liu, and F. Ricci, "Trustworthy recommender systems," *arXiv preprint arXiv:2208.06265*, pp. 1–16, 2022.

[21] M. Beladev, L. Rokach, and B. Shapira, "Recommender systems for product bundling," *Knowledge-Based Systems*, vol. 111, no. C, pp. 193–206, 2016.

[22] J. Bai, C. Zhou, J. Song, X. Qu, W. An, Z. Li, and J. Gao, "Personalized bundle list recommendation," in *WWW*, 2019, pp. 60–71.

[23] J. Hao, T. Zhao, J. Li, X. L. Dong, C. Faloutsos, Y. Sun, and W. Wang, "P-companion: A principled framework for diversified complementary product recommendation," in *CIKM*, 2020, pp. 2517–2524.

[24] L. Chen, Y. Liu, X. He, L. Gao, and Z. Zheng, "Matching user with item set: Collaborative bundle recommendation with deep attention network," in *IJCAI*, 2019, pp. 2095–2101.

[25] S. Wang, L. Cao, L. Hu, S. Berkovsky, X. Huang, L. Xiao, and W. Lu, "Hierarchical attentive transaction embedding with intra-and inter-transaction dependencies for next-item recommendation," *IEEE Intelligent Systems*, vol. 36, no. 4, pp. 56–64, 2020.

[26] M. Ludewig and D. Jannach, "Evaluation of session-based recommendation algorithms," *User Modeling and User-Adapted Interaction*, vol. 28, no. 4, pp. 331–390, 2018.

[27] S. Wang, L. Hu, Y. Wang, L. Cao, Q. Z. Sheng, and M. Orgun, "Sequential recommender systems: Challenges, progress and prospects," in *IJCAI*, 2019, pp. 6332–6338.

[28] H. Fang, G. Guo, D. Zhang, and Y. Shu, "Deep learning-based sequential recommender systems: Concepts, algorithms, and evaluations," in *Proceedings of International Conference on Web Engineering*, 2019, pp. 574–577.

[29] W. Guo, S. Wang, W. Lu, H. Wu, Q. Zhang, and Z. Shao, "Sequential dependency enhanced graph neural networks for session-based recommendations," in *DSAA*, 2021, pp. 1–10.

[30] L. Hu, L. Cao, S. Wang, G. Xu, J. Cao, and Z. Gu, "Diversifying personalized recommendation with user-session context," in *IJCAI*, 2017, pp. 1858–1864.

[31] N. Wang, S. Wang, Y. Wang, Q. Z. Sheng, and M. A. Orgun, "Exploiting intra- and inter-session dependencies for session-based recommendations," *World Wide Web Journal*, vol. 25, no. 1, pp. 425–443, 2022.

[32] H. Fang, D. Zhang, Y. Shu, and G. Guo, "Deep learning for sequential recommendation: Algorithms, influential factors, and evaluations," *ACM Transactions on Information Systems*, vol. 39, no. 1, pp. 1–42, 2020.

[33] R. Guidotti, G. Rossetti, L. Pappalardo, F. Giannotti, and D. Pedreschi, "Market basket prediction using user-centric temporal annotated recurring sequences," in *ICDM*, 2017, pp. 895–900.

[34] L. Yu, L. Sun, B. Du, C. Liu, H. Xiong, and W. Lv, "Predicting temporal sets with deep neural networks," in *SIGKDD*, 2020, pp. 1083–1091.

[35] Z. Sun, D. Yu, H. Fang, J. Yang, X. Qu, J. Zhang, and C. Geng, "Are we evaluating rigorously? benchmarking recommendation for reproducible evaluation and fair comparison," in *RecSys*, 2020, pp. 23–32.

[36] H. Ying, F. Zhuang, F. Zhang, Y. Liu, G. Xu, X. Xie, H. Xiong, and J. Wu, "Sequential recommender system based on hierarchical attention networks," in *IJCAI*, 2018, pp. 3926–3932.

[37] Y. Shen, B. Ou, and R. Li, "Mbn: Towards multi-behavior sequence modeling for next basket recommendation," *TKDD*, vol. 16, no. 5, pp. 1–23, 2022.

[38] M. Ariannezhad, S. Jullien, M. Li, M. Fang, S. Schelter, and M. de Rijke, "ReCANet: A repeat consumption-aware neural network for next basket recommendation in grocery shopping," in *SIGIR*, 2022.

[39] Z. Liu, X. Li, Z. Fan, S. Guo, K. Achan, and P. S. Yu, "Basket recommendation with multi-intent translation graph neural network," in *IEEE Big Data*, 2020, pp. 728–737.