User Modeling and User-Adapted Interaction
https://doi.org/10.1007/s11257-023-09369-8

Non-binary evaluation of next-basket food recommendation

Yue Liu[1] · Palakorn Achananuparp[1] · Ee-Peng Lim[1]

Received: 8 July 2022 / Accepted in revised form: 15 May 2023

Abstract

Next-basket recommendation (NBR) is a recommendation task that predicts a basket or a set of items a user is likely to adopt next based on his/her history of basket adoption sequences. It enables a wide range of novel applications and services from predicting next basket of items for grocery shopping to recommending food items a user is likely to consume together in the next meal. Even though much progress has been made in the algorithmic NBR research over the years, little research has been done to broaden knowledge about the evaluation of NBR methods, which is largely based on the offline evaluation experiments and binary relevance paradigm. Specifically, we argue that recommended baskets which are more similar to ground truth baskets are better recommendations than those that share little resemblance to the ground truth, and therefore, they should be granted some partial credits. Based on this notion of non-binary relevance assessment, we propose new evaluation metrics for NBR by adapting and extending similarity metrics from natural language processing (NLP) and text classification research. To validate the proposed metrics, we conducted two user studies on the next-meal food recommendation using numerous state-of-the-art NBR methods in both online and offline evaluation settings. Our findings show that the offline performance assessment based on the proposed non-binary evaluation metrics is more representative of the online evaluation performance than that of the standard evaluation metrics.

Keywords Next-basket recommendation · Food recommendation · Non-binary relevance · Evaluation metrics · User study

✉ Palakorn Achananuparp
palakorna@smu.edu.sg

Yue Liu
yueliu@smu.edu.sg

Ee-Peng Lim
eplim@smu.edu.sg

[1] Singapore Management University, Singapore, Singapore

Published online: 15 June 2023

1 Introduction

Next-basket recommendation (NBR) task is a type of sequential recommendation task (Zimdars et al. 2001) which aims to predict a collection or set of items (also known as *basket*) a user is likely to adopt at the next time step given his or her past sequence of baskets. The NBR task, first popularized by Rendle et al. (2010), has increasingly become an important area of research thanks to its prevalence in numerous real-world applications. Not only has NBR models been used predominantly in predicting a shopping basket for the customer's next purchase (Wang et al. 2015; Yu et al. 2016; Jannach and Ludewig 2017; Wan et al. 2018; Ying et al. 2018; Hu and He 2019; Le et al. 2019; Faggioli et al. 2020; Hu et al. 2020; Yu et al. 2020; Qin et al. 2021), it has also been applied to predict a variety of item sets, e.g., a set of food items for the next-meal food consumption (Bharadhwaj et al. 2018; Liu et al. 2019), a music or video playlist for the next listening/watching session (Chen et al. 2012; Hidasi et al. 2015; Kapoor et al. 2015; Jannach and Ludewig 2017; Kotzias et al. 2019; Ren et al. 2019), and a sequence of point of interests (POIs) for the next visiting period (Brilhante et al. 2013; Cheng et al. 2013; Ying et al. 2018; Kotzias et al. 2019). Compared to a conventional top-n recommendation task, which ignores sequential information in past item adoptions and only aims to infer general user preferences to recommend new items, NBR focuses on a time-specific prediction of a basket of items that a user would like to adopt next. Furthermore, the baskets may comprise both new and previously adopted items.

Although several novel NBR modeling approaches, including deep neural networks (DNNs)-based methods (Yu et al. 2016; Wan et al. 2018; Ying et al. 2018; Hu and He 2019; Le et al. 2019; Yu et al. 2020; Qin et al. 2021), have been proposed in recent years, little attention has been given to the NBR evaluation research. Following a troubling trend in the recommender systems' reproducibility (Dacrema et al. 2019), Li et al. found that limited progress has been made when comparing the offline performance of several state-of-the-art (SOTA) NBR methods and simple item popularity-based baselines (Li et al. 2021). Yet, several methodological issues in NBR evaluation have remained unexplored. Firstly, most NBR evaluations employ an offline evaluation method, commonly used in the recommender systems research at large, in which a dataset is split into the train/test partitions. The model is trained on the training set, and its prediction accuracy is evaluated against the test set as if the recommendations were shown to the users. Although offline evaluation is a valuable tool in the development of recommendation algorithms, its predictive power has been questioned (Cremonesi et al. 2012; Beel et al. 2013; Rossetti et al. 2016) and online evaluations and user studies (Shani and Gunawardana 2011; Ricci et al. 2015) still remain the most reliable methods that provide the strongest evidence of the recommender systems' performance by accounting for human factors. To our knowledge, little research has evaluated the performance of NBR methods in the online evaluations and user studies.

Secondly, offline evaluations of NBR methods typically utilize information retrieval-based metrics, such as precision, recall, and normalized discounted cumulative gain (nDCG) to measure the model accuracy (Rendle et al. 2010; Wang et al. 2015; Yu et al. 2016; Bharadhwaj et al. 2018; Liu et al. 2019; Faggioli et al. 2020; Qin et al. 2021). These metrics are built on the binary relevance assessment in which rec-

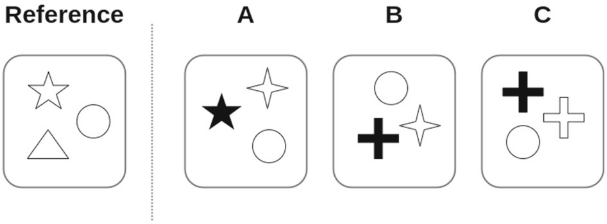


Fig. 1 Not all baskets are equal: an illustrative example

ommended items are considered relevant if they match exactly to ground truth items in the test set, whereas those that differ from the ground truth are treated as irrelevant recommendations. In the NBR context, this means that item baskets with the same number of relevant recommendations are judged to be equal in quality. However, much like Frumerman et al.'s claim (Frumerman et al. 2019) "not all rejected items in the top-*n* recommendation are equally bad," we assert that not all recommended baskets in the offline NBR evaluation should be treated the same and partial credits should be proportionately given to those which are similar to the ground truth baskets to further distinguish their quality.

To illustrate our claim, let us consider the following toy example comprising a reference basket (ground truth) of 3 items and recommended baskets A, B, and C in Fig. 1. As we can see, all three baskets contain one exact-matching item (a white circle) given the reference basket. As a result, these recommended basket will be considered of equal quality under the binary relevance assessment paradigm. On the other hand, by using a non-binary assessment that considers the similarity between shapes (i.e., circle, triangle, star, and cross) and colors (i.e., white and black) of items, basket A is the most similar to the reference basket since its two other items are more similar to the ground truth, i.e., black versus white stars and 4-pointed versus 5-pointed stars, than those of baskets B and C. Likewise, basket B is more similar to the reference basket than basket C. With this illustration, one can surmise that a user would be most satisfy with the recommended basket A and the non-binary-based evaluation will likely offer a more accurate performance assessment than the binary relevance-based evaluation.

This study aims to expand current knowledge about the evaluation of NBR methods in the food recommendation domain (Elsweiler et al. 2022) by investigating: (1) the use of various similar metrics, including those utilized in natural language processing (NLP) tasks, for the non-binary relevance assessment; (2) the performance of different NBR methods as measured by the non-binary-based metrics; (3) the effectiveness of different NBR methods in an online next-meal recommendation user study; and (4) the correspondence between user preferences for recommended item baskets and the non-binary relevance assessment of the basket quality.

To that end, we first operationalize the non-binary relevance assessment of item baskets in terms of aggregated similarity of individual items in the recommended and reference baskets. In particular, we consider two main approaches for measuring pairwise item similarity based on textual content and categorical tags. For the content-based approach, we adapt and extend several text similarity metrics widely used in NLP research (Papineni et al. 2002; Lin 2004; Zhang et al. 2020), such as machine translation, text summarization, and text generation, to measure pairwise item similarity. Next, for the tag-based approach, we propose hierarchical evaluation metrics utilizing the hierarchy of tags describing categorical information about items. Then, we incorporate a *best matching principle* to derive the basket-level assessment of non-binary relevance.

Given the content-based and hierarchical evaluation metrics for measuring pairwise item similarity, we pose our first research question as follows:

> **RQ1**: *How do different similarity metrics correspond to human similarity perception of items*?

To answer RQ1, we conduct user studies to collect human judgments of item similarity. Particularly, we are interested in two types of human perceptions: *non-personalized* and *personalized* similarity judgments. Non-personalized item similarity judgments are exercised when human annotators objectively assess the similarity between a pair of items exclusive of their own preference. In contrast, personalized item similarity judgments are employed when human annotators subjectively evaluate the similarity between a pair of items with respect to the annotators' context. The distinction between the two types of judgments is important in the food recommendation domain as the former seeks to answer the questions "how similar are food items A and B?" or "how likely is food item A a substitute for food item B in general?," whereas the latter aims to answer the question "how likely is food item A a substitute for food item B given my (the annotator's) meal context C?" which involves personal preference. We describe the two user studies for non-personalized and personalized similarity judgments in Sects. 4.1 and 4.2.2, respectively.

Next, we carry out an online next-meal recommendation user study, described in Sect. 4.2, to assess the effectiveness of various NBR methods. Fifty participants take part in an online food recommendation study in which each participant is provided with a number of algorithmically generated food item baskets for their next-meal consumption tailored to his/her past consumption data from multiple NBR algorithms, i.e., a within-subject experiment design. The participant then indicates his/her preference for each recommended item in the baskets. Specifically, the user study aims to answer the following research questions:

> **RQ2**: *How do different evaluation metrics correspond to the real users' preferences for item baskets*?

> **RQ3**: *To what extent do user preferences for item baskets differ across different NBR algorithms*?

Lastly, given the findings from RQ1 - RQ3, we conduct an offline experiment to investigate the performance of NBR methods based on the non-binary relevance assessment to answer the following research question:

> **RQ4**: *What is the offline performance of different NBR algorithms as measured by the non-binary evaluation metrics*?

Findings from this study will validate the non-binary relevance assessment paradigm in the NBR evaluations, especially the applicability of various content-based and hierarchical evaluation metrics. Furthermore, the research will provide an empirical evidence to inform the performance of several SOTA methods in the next-meal recommendation task from both the offline and online evaluation methods and bridge the gap between the binary and non-binary paradigms in the offline NBR evaluation.

Our work makes the following contributions to the NBR research area. Firstly, we propose several content-based and hierarchical evaluation metrics by adapting and extending relevant metrics from the NLP and text classification research to measure similarity between the ground truth and the recommendations at the item and basket levels. To date, we are the first to utilize such metrics in the non-binary relevance-based NBR evaluations.

Secondly, we introduce a few novel experimental protocols, including: (1) a queuing-based crowdsourcing task design for efficiently collecting pairwise item similarity judgments for the basket-level comparisons and (2) an experimental pipeline comprising online food logging, NBR algorithms, and Google Form, for conducting an online user study without the reliance on existing next-meal food recommender systems.

Thirdly, we show the validity of the NLP-based and hierarchical evaluation metrics in operationalizing the non-binary relevance assessment in the NBR evaluations. These metrics correspond more closely to human perceptions of similarity and preference than standard binary-based metrics, such as precision, recall, and nDCG. Furthermore, we uncover differences between non-binary-based metrics in their correlations with non-personalized and personalized similarity judgments and user preference judgments. Specifically, the metrics which correlate more strongly with non-personalized similarity judgments do not necessarily produce the same results with personalized similarity and preference judgments.

Lastly, our work is one of the earliest studies (Shao et al. 2021) that examine the performance of NBR methods through an online-recommendation user study. Consistent with the offline evaluation results, the participants in the online next-meal recommendation study generally prefer item baskets recommended by repeat-consumption aware NBR algorithms than those of sequential recommenders. Our findings are also in agreement with Li et al.'s analysis (Li et al. 2021) which identifies the limitations of several advanced NBR algorithms in capturing the trade-off between the repeat and explore items in the recommendations. Through both the offline and online experiments, we have also identified the non-binary-based metrics which are highly indicative of the user preferences in an online recommendation setting. These metrics will thus be useful for evaluating future online recommendation results.

The rest of the paper is organized as follows. We first survey the related work in Sect. 2. Next, we present the dataset, algorithms, and evaluation metrics used in this study, and the performance of the NBR algorithms on standard metrics as baselines in Sect. 3. Then, we describe the user studies conducted to answer the research questions in Sect. 4 and present the results of the data analysis in Sect. 5. Lastly, we discuss the limitations and future directions of our research and conclude the paper in Sects. 6 and 7, respectively.

2 Related work

We review related work from two relevant research areas: (1) offline and online evaluation of recommender systems and (2) similarity metrics and non-binary relevance, while non-accuracy-based evaluation metrics (Ge et al. 2010), such as diversity/coverage, non-redundancy, representativeness, etc., are all useful in measuring the user satisfaction of recommender systems and have been actively investigated by the recommender systems community (Ricci et al. 2015; Shani and Gunawardana 2011). Examining the relationships between those metrics and non-binary-based metrics is an interesting topic which we leave for future work.

2.1 Offline and online evaluations

Algorithmic recommender systems research has long been focusing on achieving state-of-the-art (SOTA) performance as measured by accuracy-based metrics, such as precision, recall, and nDCG, in offline evaluation settings. However, results from various studies have shown that employing the best offline algorithms does not always lead to better recommendations in a live environment (Cremonesi et al. 2012; Beel et al. 2013; Garcin et al. 2014; Rossetti et al. 2016). Since the performance of recommender systems in production is greatly affected by human factors and dynamic environments, online evaluation and user study are indispensable and complementary tools to offline evaluation.

Over the years, several online evaluations and user studies have been conducted mostly in the top-*n* recommendation evaluations under varying settings. First, a few works have investigated the consistency between results from offline and online evaluations using in live recommender systems for top-*n* movie (Cremonesi et al. 2012; Rossetti et al. 2016), news (Garcin et al. 2014), and research paper (Beel et al. 2013, 2016) recommendations, in which contradictory results from offline and online metrics have been observed. Beyond comparing the offline and online experimental results, other works (Maksai et al. 2015; Krauth et al. 2020) have examined the predictive power of accuracy and non-accuracy based offline metrics in determining online performance under various conditions. In electronic commerce (e-commerce), researchers have performed A/B tests to further validate the performance of promising recommendation methods from offline experiments in real recommender systems, including music (Domingues et al. 2013), video (Symeonidis et al. 2020), product (Kaminskas et al. 2015), and tour packages (Peska and Vojtas 2020) recommendations. When real systems are not available, user studies have been conducted to evaluate the accuracy of recommendation methods and collect data about user preferences (Yao and Harper 2018), qualitative responses, and feedback from real users (Braunhofer et al. 2013; Kamehkhosh and Jannach 2017) or domain experts (Messina et al. 2019; Färber and Sampath 2020).

In food recommender systems, most performance evaluations have been done almost exclusively in offline experiments (Trattner and Elsweiler 2017). Online evaluations and user studies in the food recommendation research have been conducted in the past few years (Elsweiler et al. 2022), mostly in the top-*n* cooking recipe recommendation domain (Ge et al. 2015; Massimo et al. 2017; Musto et al. 2020; Trattner and Jannach 2020; Hauptmann et al. 2021). While most food and recipe recommendation studies have been conducted with study participants in short single experimental sessions or through online crowdsourcing platforms (Achananuparp and Weber 2016; Musto et al. 2020; Trattner and Jannach 2020), some studies, especially on the health-aware recipe recommendation, have employed a more rigorous controlled experiment design which took place over several weeks (Achananuparp et al. 2018; Hauptmann et al. 2021).

Lastly, Shao et al. (2021) recently conducted an online user study to evaluate course recommender systems with college students. As their multi-semester course recommendation is formulated as the NBR problem, the user study is considered one of the earliest to be performed in the context of NBR evaluation.

2.2 Similarity metrics and non-binary relevance

Measuring text similarity has a long history in NLP and information retrieval (Robertson et al. 1995). More recently, much effort has been focusing on assessing the similarity of sentences or short texts. Early methods are based on word overlap (Metzler et al. 2005) and bag-of-words model incorporating external knowledge sources (Li et al. 2006; Achananuparp et al. 2008, 2009). Word or n-gram overlap-based methods (Papineni et al. 2002; Lin 2004) are commonly used in NLP evaluations thanks to their computational efficiency and strong correlation with human perception of similarity. Over the years, a learning-based approach, including deep-learning-based (He et al. 2015; Mueller and Thyagarajan 2016; Peng et al. 2020), word moving distance-based (Kusner et al. 2015; Huang et al. 2016), and embeddings-based methods (Le and Mikolov 2014; Kenter and De Rijke 2015; Kiros et al. 2015; Arora et al. 2017; Zhang et al. 2020; Sellam et al. 2020; Sun et al. 2022), has gained much attention due to an effective use of growing number of large datasets to pre-compute/pre-train models.

In the recommender systems research, computing similarity of items or users is a long-standing task at the core of several recommender systems' mechanics. Firstly, item-based collaborative filtering (CF) recommender systems (Sarwar et al. 2001) normally compute item similarity from the ratings or interactions data when performing a k-nearest neighbors (k-NN) algorithm to predict item ratings. Secondly, content-based (CB) recommender systems (Lops et al. 2011) typically utilize the TF-IDF weighted vector-space model and other information retrieval methods for item similarity computation. The CB similarity has also been incorporated into CF recommender systems to improve the recommendation performance (Melville et al. 2002). Next, past research has shown strong correlation between content-based similarity scores and human judgments of item similarity in the respective domains, such as similar movies recommendation (Colucci et al. 2016; Yao and Harper 2018; Trattner and Jannach 2020) and similar cooking recipes recommendation (Trattner and Jannach 2020).

Motivated by the assumption that some rejected or non-interacted items in the recommendations are more valuable than others (Lacic et al. 2019; Frumerman et al. 2019; Sánchez and Bellogín 2019), recent research has explored how non-binary relevance can be incorporated into the offline evaluation of top-n recommendations. Specifically, partially relevant items are defined as those with non-zero similarity scores, compared to ground truth. For item similarity computation, a few non-binary relevance-based evaluation metrics have been proposed, including CB (Lacic et al. 2019; Frumerman et al. 2019; Sánchez and Bellogín 2019) and CF-based (Frumerman et al. 2019) approaches. By analyzing the user-item interaction data (e.g., item clicks), the CB similarity metrics has been shown to have a stronger correlation with the number of item clicks than the CF-based similarity and the standard precision metrics (Frumerman et al. 2019). Two recent CB item similarity methods include a *doc2vec* (Le and Mikolov 2014) embedding-based cosine similarity method (Lacic et al. 2019) and attribute-based item similarity methods (Frumerman et al. 2019; Sánchez and Bellogín 2019) incorporating the exact matching of item attributes (Brilhante et al. 2013; He et al. 2017), such as job titles, movie genre, venue categories, etc.

2.3 Comparison with previous work

Our work shares some commonalities and differences to previous work on the non-binary relevance assessment in the recommender systems evaluation, especially Lacic et al. (2019); Frumerman et al. (2019) and Sánchez and Bellogín (2019). Firstly, our research is motivated by the similar assumption as theirs in that partially relevant recommendations could still be useful in the offline evaluations. We, however, focus on the next-basket recommendation evaluation, whereas prior work examined the top-n recommendation evaluation. Thus, the units of recommendation are different, i.e., baskets versus items.

Next, our proposed non-binary-based metrics are directly built on the previous work, incorporating content-based approaches such as embedding-based (Lacic et al. 2019) and item-attribute-based item similarity metrics. Unlike Lacic et al.'s simple adoption of the *doc2vec* model for tag similarity, we adapt and extend various NLP-based metrics, such as BLEU (Papineni et al. 2002), ROUGE (Lin 2004), and BERTScore (Zhang et al. 2020), for basket similarity computation.

Our proposed hierarchical evaluation metric is conceptually similar to the attribute-based metrics in Frumerman et al. and Sánchez and Bellogín as both methods proportionately compare the overlap between item attributes or categories. Nevertheless, ours does not require specific item attributes to be manually selected and weighted for the similarity computation. Furthermore, we also explore a hybrid approach which combines NLP-based and hierarchical evaluation metrics in determining item similarity. Lastly, none of the previous work has directly evaluated their metrics against human judgments of similarity and preference for recommendations, which is crucial in gauging their validity.

In terms of online evaluation methods, ours and Shao et al. (2021) are among the earliest studies which utilize both offline and online experiments to validate NBR algorithms. Although Shao et al. (2021) has recently conducted an online user study for NBR, their course recommendation domain is drastically different from most NBR problems in grocery shopping (Rendle et al. 2010; Wang et al. 2015), food consumption (Bharadhwaj et al. 2018; Liu et al. 2019), and music listening (Chen et al. 2012; Hidasi

Table 1 Dataset statistics

#users	#items	#transactions	density	#baskets	#items per user	basket size	%repeat consumption
6,916	47,789	2,260,319	0.23%	414,874	107.68 ± 79.8	5.45 ± 3.39	55.69% ± 18.77%

et al. 2015) which are commonly characterized by the dynamics of repeat and novel consumptions.

3 Dataset, algorithms, and evaluation metrics

We begin by introducing the materials used in this study, including the dataset, algorithms, and evaluation metrics, in Sects. 3.1, 3.2, and 3.3, respectively.

3.1 Dataset

In this study, we utilize a public food diary dataset **MyFitnessPal** (MFP) (Weber and Achananuparp 2016), consisting of 587K food diaries logged by 9.9K users over a 6-month period. Each food diary can been seen as a basket of food items representing daily food intake of each user. Each food item consists of a textual description and is automatically annotated with one or more categorical tags from a tag hierarchy using a keyword matching method (Weber and Achananuparp 2016). For example, the annotated hierarchical tag, *fruit* → *tropical* → *banana*, shows that the food item is given *fruit* as the first-level tag, followed by *tropical* as the second-level tag, and followed by *banana* as the third-level tag. In total, there are 19 first-level tags, 85 s-level tags, and 1,263 third-level tags in the hierarchy. Since an item can be associated with multiple tags at the same level of tag hierarchy, the dataset contains 17K tag combinations for the 47K items where a tag combination is a unique set of all tags assigned to an item. For example, the following items '*classic tuna sandwich*' and '*sandwich with tuna spread*' share the same tag combination {*staple* → *wheat* → *bread*, *meat* → *fish* → *tuna*}. Given its textual contents and large hierarchy of tags, MFP is an ideal dataset for our study.

We followed the same data cleaning procedures used in Liu et al. (2019). Specifically, we performed p-core filtering by recursively removing: (1) items that were adopted by less than 20 users; (2) users who adopted less than 5 remaining items; and (3) users who recorded no more than 2 days of food diaries. After data preprocessing, the dataset statistics, including mean ± standard deviation, are described in Table 1. As we can see, the dataset is highly sparse and contains a large degree of repeat consumption, which are common characteristics of many NBR datasets (Wan et al. 2018; Kotzias et al. 2019; Li et al. 2021). Repeat consumption occurs when a user adopted the same item more than once.

Table 2 List of symbols

Symbols	Description
U	set of users, $\{u_1, u_2, ..., u_i, ..., u_{\|U\|}\}$
V	set of items, $\{v_1, v_2, ..., v_j, ..., v_{\|V\|}\}$
$G(i)$	ground truth basket of u_i
$RecList(i, k)$	top-k recommended items for u_i
$rank(i, j)$	rank of v_j in the recommended basket for u_i
γ^n	hyperparameter weight of n-gram precision in BLEU-N
$p_n(j'\|j)$	n-gram precision of items j' with respect to j
$T_n(j)$	list of n-gram tokens extracted from v_j's content
w_j	tokenized word vectors of v_j's content
C_j	set of tags associated with v_j
λ_t	importance of tag c_t
h	weight ratio of a child node to its parent node

3.2 Recommendation algorithms

The next-basket (NBR) food recommendation task involves predicting a basket of food items (also referred to as meals) the user is likely to consume next given his/her past food intake data. In this study, a basket consists of food items consumed in a whole day across all meal occasions (e.g., breakfast, lunch, etc.). To generate next-meal recommendations for the user studies, we select a few state-of-the-art (SOTA) NBR algorithms as well as commonly used baseline methods in related tasks, such as session-based recommendation and top-n recommendation. These algorithms cover four diverse modeling approaches, sufficiently reflecting the NBR research landscape: (1) naive non-personalized baselines, (2) repeat-consumption aware algorithms incorporating the dynamics of repeat and novel item adoption, (3) standard latent factor-based item recommendation algorithms, and (4) sequential basket recommendation algorithms. All algorithms produce top-k recommended items as baskets. For convenience, all mathematical symbols and notations used in this section as well as subsequent sections are presented in Table 2.

Naive non-personalized algorithms. Two naive non-personalized baselines that utilize simple heuristics include:

- **Random**: A random baseline where each item v_j is assigned a random score for each user u_i. The k items with highest scores will be returned as the recommended basket.
- **Global**: Global popularity is a commonly used naive baseline for top-k recommendation tasks (Rendle et al. 2009) and has been shown to perform well on some datasets (Dacrema et al. 2019). Each item v_j is assigned a score proportional to its global adoption frequency n_j in the training (and validation) set. The recommended basket consists of k items with the highest frequencies.

Repeat consumption-aware algorithms. Four algorithms that specifically model the dynamics of repeat novel consumption of individual users over time are:

- **Personal**: Personal popularity is a naive personalized algorithm which simply recommends for a user u_i k items with the highest adoption frequencies in u_i's past history, i.e., a repeat consumption only recommendation. Similar to Global, the Personal baseline can perform very competitively in many item recommendation datasets (Dacrema et al. 2019).
- **Mixture**: Multinomial mixture model (Kotzias et al. 2019) is an exploration–exploitation-based mixture model that predicts the likelihood user u_i adopts item v_j by balancing the trade-off between the novel and repeat consumptions. It substantially outperformed the matrix factorization and global popularity baselines on the online forum posts, songs, and check-ins datasets (Kotzias et al. 2019).
- **MixtureTW**: Time-weighted multinomial mixture model (Liu et al. 2019) is a simple extension to Mixture (Kotzias et al. 2019) in which adoption frequency is discounted with an exponential time decay. It achieved the state-of-the art (SOTA) performance in next-meal recommendation on the MFP dataset, outperforming the original Mixture, matrix factorization-based, and popularity-based baselines (Liu et al. 2019).
- **adaLoyal**: Triple2vec + adaLoyal (Wan et al. 2018) is an embedding-based representation learning method and a recommendation algorithm that balance the repeat consumption with user preferences by explicitly modeling items complementarity, compatibility, and loyalty. It attained the SOTA performance on the grocery shopping datasets over the embedding-based and popularity-based baselines at the time of its publication (Wan et al. 2018).

Latent factor-based algorithms. We include four latent factor-based algorithms which learn latent factors of user-item interactions to infer user preferences for general top-*n* recommendations:

- **NMF**: Non-negative matrix factorization (Lee and Seung 2000) is a popular latent factor-based algorithm commonly recognized as a strong baseline in a variety of recommendation tasks.
- **BPR-MF**: Bayesian personalized ranking (Rendle et al. 2010) is an extension of NMF that uses pairwise ranking loss shown to be especially effective on recommendation tasks with implicit feedback data.
- **WRMF**: Weighted regularized matrix factorization (Hu et al. 2008) is a matrix factorization model that assigns weights on consumption frequency shown to be highly effective on count data.
- **LDA**: Latent Dirichlet allocation is a well-known probabilistic topic model used as a competitive baseline in several recommendation tasks (Gopalan et al. 2015; Trattner and Elsweiler 2017; Kotzias et al. 2019).

Sequential recommendation algorithms. Lastly, we chose two sequential recommender-based algorithms which directly model a sequence of past adoptions/interactions between users and items to generate personalized recommendation lists.

- **FPMC**: Factorized personalized Markov Chains (Rendle et al. 2010) is a classical NBR method which combines both Markov chains and matrix factorization to capture both short-term item-to-item transitions and long-term preferences in the basket sequence data, respectively. It is generally recognized as a highly competitive baseline in various settings.

- **SASRec**: Self-attentive sequential recommendation model (Kang and McAuley 2018) is a self-attention (Vaswani et al. 2017)-based DNN model which is shown to be highly effective in a session-based recommendation task. It sets a new SOTA performance on the online shopping, online games, and movies recommendation datasets, outperforming all other methods including DNN-based sequential recommenders, BPR-MF, FPMC, and popularity-based baseline (Kang and McAuley 2018). To adapt the original SASRec to the NBR task, we applied max pooling operations, similar to Wang et al. (2015), to create a basket representation from item representations.

As we can see, the characteristics of item baskets recommended by different algorithms may vary depending on how the historical baskets data are modeled. For example, baskets generated by conventional top-k algorithms, such as Global, Personal, etc., consist of items selected individually and independently of one another, whereas baskets generated by more sophisticated algorithms like adaLoyal contain items with specific relationships to each other, e.g., complementarity.

3.3 Evaluation metrics

We formally define all evaluation metrics used in this study, including (1) standard binary-based metrics, (2) non-binary content-based metrics, and (3) non-binary hierarchical evaluation metrics.

3.3.1 Standard binary-based evaluation metrics

Standard information retrieval-based metrics, such as recall, precision, and normalized discounted cumulative gain (nDCG), are commonly used to evaluate algorithmic accuracy of the next-basket recommendation task based on binary relevance between the top-k recommended items and ground truth items where $k \in [1, \infty)$. The commonly used top-k metrics include recall@k, precision@k, and nDCG@k.

Firstly, **recall@k** measures the proportion of ground truth next-basket items $G(i)$ for user u_i correctly recommended among the top-k recommendation items $RecList(i, k)$ as shown in Eq. 1.

$$\text{Recall@k} = \frac{1}{|U|} \sum_{u_i \in U} \frac{|G(i) \cap RecList(i, k)|}{|G(i)|} \tag{1}$$

Secondly, **precision@k** measures the proportion of correctly recommended ground truth items among the top-k recommended items as defined in Eq. 2.

$$\text{Precision@k} = \frac{1}{|U|} \sum_{u_i \in U} \frac{|G(i) \cap RecList(i, k)|}{k} \tag{2}$$

Lastly, **nDCG@k** is defined in Eq. 3 as a discounted cumulative gain (DCG) of items in $RecList(i, k)$ normalized by the ideal DCG (IDCG), which is simply the

DCG measure of the best ranking result (Järvelin and Kekäläinen 2002). nDCG is found to have higher discriminative power than other metrics in evaluating top-n recommendation algorithms (Valcarce et al. 2018).

$$\mathrm{nDCG@k} = \frac{1}{|U|} \sum_{u_i \in U} \mathrm{nDCG@k}(i) \tag{3}$$

where

$$\mathrm{nDCG@k}(i) = \frac{\mathrm{DCG@k}(i)}{\mathrm{IDCG@k}(i)} \tag{4}$$

$$\mathrm{DCG@k}(i) = \sum_{v_j \in G(i) \cap RecList(i,k)} \frac{1}{\log_2(rank(i,j)+1)} \tag{5}$$

$rank(i, j)$ refers to the rank of item v_j in $RecList(i, k)$. The values of recall@k, precision@k, and nDCG@k range from 0 to 1. The higher the score, the better the recommendation accuracy.

3.3.2 Non-binary content-based evaluation metrics

In cases where textual contents of items (e.g., item name, description, etc.) are available, non-binary evaluation metrics can be defined based on the similarity between items in the recommended and ground truth baskets. Following a standard notion of relevance in information retrieval, the similarity scores can be used to represent non-binary relevance between items. We consider both **n-gram** and **embedding-based** approaches define content-based item similarity.

As any two items can have nonzero similarity score, it is important for a non-binary content-based evaluation metric to find the best matched ground truth basket item for each item in the recommended basket and ignore the non-best matched ones. We call this the *best matching principle* and apply it to all the non-binary content-based evaluation metrics. This principle is, however, only fair when we impose the same basket size restriction (i.e., top-**k** recommended items) to all the recommendation models, as performed in this work. In the following, we describe the n-gram and embedding-based metrics.

N-gram-based metrics. Firstly, we utilize simple n-gram-based metrics, widely used in various natural language processing (NLP) evaluations (e.g., machine translation, text summarization, and question answering). These metrics include **Bilingual Evaluation Understudy Score** (**BLEU**) (Papineni et al. 2002) and **Recall-Oriented Understudy for Gisting Evaluation** (**ROUGE**) (Lin 2004). These metrics have traditionally been shown to correlate well with human judgements; however, they have not been utilized in the next-basket food recommendation evaluation.

We define BLEU-N@k for the predicted basket with top-k recommended items as shown in Eq. 6 where N is the length of n-gram of item content used in matching recommended and ground truth items.

$$\text{BLEU-N@k} = \frac{1}{|U|} \sum_{u_i \in U} Avg_{v_{j'} \in RecList(i,k)} \max_{v_j \in G(i)} \text{BLEU-N}(j'|j) \tag{6}$$

where

$$\text{BLEU-N}(j'|j) = \exp(\sum_{n=1}^{N} \gamma^n \log p_n(j'|j)) \tag{7}$$

$$p_n(j'|j) = \frac{|T_n(j') \bigcap T_n(j)|}{|T_n(j')|} \tag{8}$$

We ignore the brevity penalty in Eq. 7, typically used in the original formulation (Papineni et al. 2002) when evaluating machine translation models, as there is no reason to penalize short *n*-grams (e.g., number of words) when comparing item contents. By default, the BLEU metric calculates the cumulative 4-gram BLEU score by geometric mean (i.e., N=4), with uniform weight $\gamma^n = 1/N$ for *n*-gram precision in BLEU-N. However, if there is no overlap between predicted and ground truth items' content 4-grams, BLEU-4 score will be zero. In our context, since the length of item content tokens to be matched (e.g., word tokens in item names and descriptions) is generally short, we only consider N=1 ($\gamma^1 = 1$) and N=2 ($\gamma^1 = \gamma^2 = 0.5$) variants, and correspondingly BLEU-1@k and BLEU-2@k, respectively. In this metric definition, BLEU-N$(j'|j)$ returns matching score of a predicted item $v_{j'}$ given a ground truth item v_j. Based on the best matching principle, we select the best matching ground truth item for each predicted item.

The definition of **ROUGE-N@k** is shown in Eq. 9. In this work, we use N=1, 2, and L variants of ROUGE. While ROUGE-1 and ROUGE-2 are about evaluating content recall at the unigram and bigram levels, the ROUGE-L variant measures the recall of longest common subsequence between the textual contents of a recommended item with respect to a ground truth item. The values of BLEU@k and ROUGE@k range from 0 to 1. The higher the score, the more similar the items.

$$\text{ROUGE-N@k} = \frac{1}{|U|} \sum_{u_i \in U} Avg_{v_{j'} \in RecList(i,k)} \max_{v_j \in G(i)} \text{ROUGE-N}(j'|j) \tag{9}$$

where

$$\text{ROUGE-N}(j'|j) = \frac{|T_n(j') \bigcap T_n(j)|}{|T_n(j)|} \tag{10}$$

Embedding-based metrics. Next, we explore an embedding-based metric **BERTScore** (Zhang et al. 2020), which considers item's content semantics as defined by their word embeddings from a large pre-trained language model BERT (Devlin et al. 2018), as another content-based evaluation metric. One major advantage of BERTscores over the *n*-gram-based metrics is that it is able to measure semantic similarity between items even if they share no common tokens as each token will be represented by its word embedding. It has recently been shown that BERTScore correlates better with human judgements than BLEU and ROUGE in many NLP tasks, but its effectiveness has not been investigated in the next-basket food recommendation domain.

Therefore, we propose the following top-*k* evaluation metrics: $\mathbf{P_{BERT}@k}$, $\mathbf{R_{BERT}@k}$, and $\mathbf{F1_{BERT}@k}$ as defined in Eqs. 11, 12, and 13, respectively. Their values are from -1 (most dissimilar) to 1 (most similar). In this work, we used BERTScore with a default *RoBERTa* (Liu et al. 2019) *large* model for English language.

$$\mathrm{P}_{BERT}@\mathrm{k} = \frac{1}{|U|} \sum_{u_i \in U} Avg_{v_{j'} \in \mathrm{RecList}(i,k)} \max_{v_j \in \mathrm{G}(i)} \mathrm{P_{BERT}}(j'|j) \tag{11}$$

$$\mathrm{R}_{BERT}@\mathrm{k} = \frac{1}{|U|} \sum_{u_i \in U} Avg_{v_{j'} \in \mathrm{RecList}(i,k)} \max_{v_j \in \mathrm{G}(i)} \mathrm{R_{BERT}}(j'|j) \tag{12}$$

$$\mathrm{F1}_{BERT}@\mathrm{k} = \frac{1}{|U|} \sum_{u_i \in U} Avg_{v_{j'} \in \mathrm{RecList}(i,k)} \max_{v_j \in \mathrm{G}(i)} \mathrm{F1_{BERT}}(j'|j) \tag{13}$$

where

$$\mathrm{P}\ _{BERT}(j'|j) = \frac{1}{|w'_j|} \sum_{w'_{jl} \in w'_j} \max_{w_{jk} \in w_j} w_{jk}^{\intercal} \cdot w'_{jl} \tag{14}$$

$$\mathrm{R}\ _{BERT}(j'|j) = \frac{1}{|w_j|} \sum_{w_{jk} \in w_j} \max_{w'_{jl} \in w'_j} w_{jk}^{\intercal} \cdot w'_{jl} \tag{15}$$

$$\mathrm{F1}\ _{BERT}(j'|j) = 2\frac{P_{BERT}(j'|j) \cdot R_{BERT}(j'|j)}{P_{BERT}(j'|j) + R_{BERT}(j'|j)} \tag{16}$$

3.3.3 Non-binary hierarchical evaluation metrics

In addition to the non-binary content-based evaluation metrics, we propose new hierarchical evaluation metrics that exploit the hierarchy of item attributes, categories, or tags to determine the similarity between item baskets in the next-basket food recommendation task. Specifically, the proposed hierarchical evaluation metrics are inspired by the hierarchical F1 measure in the text classification evaluation (Kiritchenko et al. 2005).

We first define a recall-oriented hierarchical matching function of a recommended item $v_{j'}$ given a ground truth item v_j. Let the set of tags of item v_j be C_j, and the importance of a tag $c_t \in C_j$ be λ_t. The hierarchical matching is:

$$\text{hMatch-}\lambda(j'|j) = \frac{\sum_{c_t \in (C_{j'} \cap C_j)} \lambda_t}{\sum_{c_t \in C_j} \lambda_t} \tag{17}$$

Various weighting schemes for the parameter λ_t can be incorporated. In this work, we define a weighting scheme based on the tag's level in the hierarchy, e.g., $\lambda_t = 1$ if

t is at the root level, and $\lambda_t = h \cdot \lambda_{t_p}$ if t is the child of tag t_p. We experiment with $h \in \{1, 2\}$ in this study and hence the variants **hMatch-1** and **hMatch-2**, respectively. Alternatively, we also explore an inverse document frequency (IDF)-like scheme which assigns the importance of each tag t based on its rarity. This variant of hMatch-λ is denoted by **hMatch-IDF**.

We then propose **hierarchical precision** (**hP-˘@k**) and **hierarchical recall** (**hR-˘@k**) for predicted baskets with top-k recommended items for all users:

$$\text{hP-}\lambda@k = \frac{1}{|U|}\sum_{u_i \in U} Avg_{v_{j'} \in \mathrm{RecList}(i,k)} \max_{v_j \in \mathrm{G}(i)} \text{hMatch-}\lambda(j'|j) \tag{18}$$

$$\text{hR-}\lambda@k = \frac{1}{|U|}\sum_{u_i \in U} Avg_{v_j \in \mathrm{G}(i)} \max_{v_{j'} \in \mathrm{RecList}(i,k)} \text{hMatch-}\lambda(j'|j) \tag{19}$$

Again, we apply the best matching principle to hP-λ@k to allow each predicted basket item to be matched with the most similar ground truth basket item, and to hR-λ@k to allow each ground truth basket item to be matched with the most similar predicted basket item, respectively. The values of hP-λ@k and hR-λ@k range from 0 to 1. The higher the score, the more similar the items.

3.3.4 Non-binary hybrid hierarchical and content-based metrics

Next, we propose a hybrid extension of the hierarchical evaluation metric which incorporates the tag-level content similarity in the evaluation. We first introduce a recall-oriented hierarchical matching function with tag-level similarity **hMatch**$_{sim}$ as an extension to hMatch. Its generic form is defined as follows:

$$\text{hMatch}_{\text{sim}}-\lambda(j'|j) = \frac{\sum_{c_t \in C_j} \lambda_t \cdot \max_{c_s \in C_{j'}} sim(c_t, c_s)}{\sum_{c_t \in C_j} \lambda_t} \tag{20}$$

Similarly, the corresponding hierarchical precision metric with tag-level similarity and hierarchical recall metric with tag-level similarity are defined as:

$$\text{hP}_{\text{sim}}-\lambda@k = \frac{1}{|U|}\sum_{u_i \in U} Avg_{j' \in \mathrm{RecList}(i,k)} \max_{v_j \in \mathrm{G}(i)} \text{hMatch}_{\text{sim}}-\lambda(j'|j) \tag{21}$$

$$\text{hR}_{\text{sim}}-\lambda@k = \frac{1}{|U|}\sum_{u_i \in U} Avg_{v_j \in \mathrm{G}(i)} \max_{v_{j'} \in \mathrm{RecList}(i,k)} \text{hMatch}_{\text{sim}}-\lambda(j'|j) \tag{22}$$

Different content-based similarity functions can be used when realizing hMatch$_{sim}$. In this work, we utilize F1$_{BERT}$ for measuring the content similarity of two tags t and s, $sim(c_t, c_s)$, due to its overall effectiveness in NLP tasks (Zhang et al. 2020). In **hP**$_{\text{sim}} - \lambda@k$, the best matching principle allows each item in the predicted basket to be matched with the most similar ground truth basket item; in **hR**$_{\text{sim}} - \lambda@k$, the best matching principle allows each ground truth basket item to be matched with the most similar predicted basket item, respectively.

Table 3 Performance of different algorithms with standard metrics on a hold-out MFP test set. Best results are in boldface

Method	Precision@10	Recall@10	nDCG@10
Random	0.000	0.000	0.000
Global	0.031	0.073	0.068
Personal	0.134	0.336	0.308
Mixture	0.135	0.339	0.311
MixtureTW	**0.165**	**0.412**	**0.377**
adaLoyal	0.127	0.317	0.279
NMF	0.061	0.155	0.166
BPR-MF	0.062	0.149	0.104
WRMF	0.054	0.128	0.106
LDA	0.031	0.077	0.070
FPMC	0.143	0.324	0.293
SASRec	0.113	0.285	0.268

3.4 Training the recommendation models

We used the MFP dataset and the algorithms introduced in Sects. 3.1 and 3.2, respectively, to train the next-basket recommendation models which were subsequently employed in the user studies. We applied the following rules to split the MFP dataset into train, validation, and test sets: (1) for users who have more than one baskets, their most recent basket is used for testing; (2) for users who have more than two baskets, their second-to-last basket is used for validation; and (3) the remaining baskets are used for training. All models were trained to generate a top-k ranked list of unique items as an item basket for recommendation based on a set of all items in the training and validation set. Item baskets are treated as a set, i.e., items only appear once per basket. Based on the dataset characteristics, we set $k = 10$. The hyperparameters were tuned by optimizing the nDCG@10 metric in the validation set and the optimal settings for each model are as follows:

- MixtureTW: Decay weight = 0.9
- adaLoyal: Number of latent factors = 500, initial product loyalty = 0.9
- NMF: Number of latent factors = 100
- BPR-MF: Number of latent factors = 500, number of epochs = 100
- WRMF: Number of latent factors = 50, number of epochs = 150, L2-norm regularization coefficient = 0.01
- LDA: Number of latent factors = 50
- FPMC: Number of latent factors = 500, L2-norm regularization coefficient = 0.01, learning rate = 0.01, number of epochs = 2
- SASRec: Default values per (Kang and McAuley 2018), e.g., number of hidden units = 50, batch size = 128, learning rate = 0.001, number of epochs = 201, drop rate = 0.5
- Random, Global, Personal, and Mixture have no hyperparameters.

The performance scores were reported on the hold-out test set as shown in Table 3. MixtureTW is the best overall performer on the precision, recall, and nDCG metrics.

Table 4 Statistics of the user studies

	Study I	Study II (p1)	Study II (p2)
Number of workers/participants	241	48	48
Number of item pairs/items judgments	7,240	2,400	5,458

Moreover, repeat consumption-aware algorithms, including a naive Personal baseline, tend to perform better than the other algorithms across all metrics. The scores for the Random baseline are zero which is to be expected. Interestingly, the more sophisticated algorithms such as adaLoyal, FPMC, and SASRec do not outperform the Personal baseline in most cases, except for FPMC which outperforms Personal and is the second best algorithm on the precision metric.

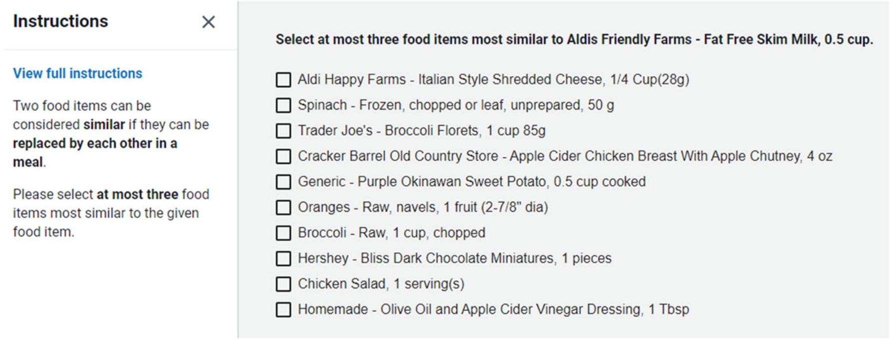
4 User studies

To answer the research questions posed in Sect. 1, we conducted user studies I and II (parts 1 and 2), described in Sects. 4.1 and 4.2, respectively. The first study aims to collect non-personalized pairwise similarity judgments, whereas the second study aims to collect user preference (part 1) and personalized pairwise similarity (part 2) judgments. For ease of referencing, we summarize basic statistics of the two user studies in Table 4.

4.1 Study I: basket-level item similarity survey

In study I, we formulated an item similarity evaluation to investigate how well different similarity scores, computed by the non-binary-based metrics, correspond to human similarity judgments. We seek to obtain human similarity judgments for a k-item recommended basket given an m-item ground truth basket from workers of an online crowdsourcing platform Amazon Mechanical Turk (AMT).

This item basket similarity judgment task is not trivial nor easy to perform as annotators are likely to experience an information overload if they were asked to exhaustively compare all $m \times k$ item pairs, adversely affecting the quality of their decisions (Malhotra 1982). To overcome the problem, we propose a novel queuing-based task design for efficient pairwise comparisons by decomposing basket comparison tasks into smaller chunks and intervals (Jones and Kelly 2018). Specifically, each annotator only needs to judge k item pairs at a time instead of $m \times k$ pairs. Then, pairwise judgments from multiple annotators are aggregated to derive human similarity judgments for all item pairs in the baskets. Figure 2 displays a screenshot of the proposed human intelligent task (HIT) conducted on the AMT platform. As we can see, each task consists of a reference item v_j (shown in boldface at the top of the screen) and a set of 10 candidate items denoted by F_j. Workers were explicitly instructed that two items are similar if they could replace each other in the same meal context (i.e., breakfast, lunch, or


Instructions

View full instructions

Two food items can be considered **similar** if they can be **replaced by each other in a meal**.

Please select **at most three** food items most similar to the given food item.

Select at most three food items most similar to Aldis Friendly Farms - Fat Free Skim Milk, 0.5 cup.

- □ Aldi Happy Farms - Italian Style Shredded Cheese, 1/4 Cup(28g)
- □ Spinach - Frozen, chopped or leaf, unprepared, 50 g
- □ Trader Joe's - Broccoli Florets, 1 cup 85g
- □ Cracker Barrel Old Country Store - Apple Cider Chicken Breast With Apple Chutney, 4 oz
- □ Generic - Purple Okinawan Sweet Potato, 0.5 cup cooked
- □ Oranges - Raw, navels, 1 fruit (2-7/8" dia)
- □ Broccoli - Raw, 1 cup, chopped
- □ Hershey - Bliss Dark Chocolate Miniatures, 1 pieces
- □ Chicken Salad, 1 serving(s)
- □ Homemade - Olive Oil and Apple Cider Vinegar Dressing, 1 Tbsp

Fig. 2 Basket-level item similarity survey

dinner). Then, their task is to select up to three candidate items in F_j that are most similar to the reference item v_j.

We generated the data for the basket-level similarity HITs as follows. Given the MFP dataset, a subset of 40 MFP users were randomly chosen. For each user u_i, the items from the last basket consumed by him/her denoted by $G(i)$ were used as the reference items. Then, we combined the top-10 recommended items returned by recommendation algorithms for the last basket into a positive candidate item set $PosList(i)$. That is, $PosList(i) = \cup_{algo \in ASet, k=10} \text{RecList}_{algo}(i, k)$ where $ASet =$ {Personal, MixtureTW, NMF, FPMC, SASRec}. An algorithm in $ASet$ was selected for each representative approach to the next-basket recommendation task and its relative performance in the offline evaluation as outlined in Sect. 3.2. Then, each item in $G(i)$ was later used as a reference item v_j, and F_j was assigned 8 candidate items randomly selected from $PosList(i)$ and 2 candidate items randomly selected from $V - PosList(i)$. Note that v_j may also appear among the candidate items in F_j. A HIT is formed by v_j and F_j and is assigned to three AMT workers. As a result, 724 HITs were generated and used for collecting human judgments of 7,240 item pairs. Out of these item pairs, 5,982 are unique.

For each reference item v_j, we use $n(v_{j'}, v_j)$ and $vote(v_{j'}, v_j)$ to denote the number of workers assigned to judge if $v_{j'} \in F_j$ is similar to v_j and the number of them voting $v_{j'}$ to be similar to v_j, respectively. Majority voting strategy was used to obtain the final human perceived similarity score:

$$Sim_{human}(v_{j'}|v_j) = \begin{cases} 1, & \text{if } \frac{vote(v_{j'}, v_j)}{n(v_{j'}, v_j)} > 0.5 \\ 0, & \text{otherwise.} \end{cases} \tag{23}$$

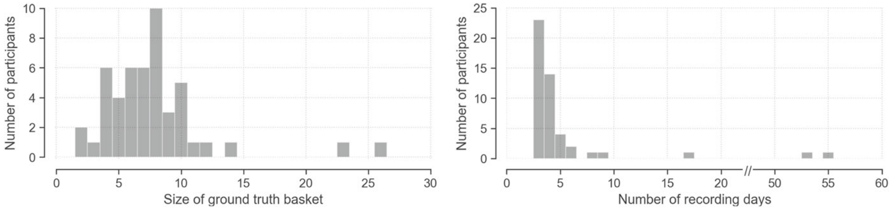
After aggregating all crowdsourced judgment data, 90.66% (5,423 out of 5,982) of rated item pairs have Sim_{human}=0, whereas the remaining 9.34% (559 pairs) have Sim_{human}=1. Among the 61 rated pairs that contain identical items, only one pair has Sim_{human}=0. This shows that the workers are fairly attentive in performing the tasks. It is worth noting that inter-rater reliability of the basket-based similarity task is low according to the Krippendorf's alpha score of 0.256, suggesting that identifying similar food items is highly subjective. Nevertheless, the human similarity judgments are still useful in identifying the best similarity functions and the corresponding non-binary evaluation metrics.

4.2 Study II: online next-basket recommendation evaluation

Study II was formulated as an online recommendation evaluation and conducted in a two-part online personalized survey: the next-basket recommendation survey in part 1 and the personalized basket-level similarity survey in part 2. In the next-basket recommendation survey, participants were asked to evaluate *their own preference* for the recommended items in a food basket generated by different recommendation algorithms, i.e., a within-subject design. After completing part 1, participants were then asked to judge the similarity/substitutability between the actual (i.e., ground truth) and recommended baskets of items. Unlike in study I, each item pair in study II's personalized basket-level similarity survey was rated by one annotator whose food diary data were used to construct the item pairs in the survey. It should be noted that the two-part survey was specifically structured to encourage participants to independently use his/her own decision criteria in the preference judgments in part 1 without being inadvertently primed or influenced by the item similarity criteria, which they were asked to exercise later in part 2.

As we did not have access to a live next-basket food recommender system with real and active users, we implemented an experimental pipeline that utilizes the public MFP dataset (described in Sect. 3.1), the online food logging tool MyFitnessPal, and the Amazon Mechanical Turk (AMT) to conduct the online user study. We aimed to enroll 50 qualified participants, a suitable sample size for experimental research (Delice 2010), from a pool of available AMT workers. The recruitment was done on the AMT platform and was restricted to workers who resided in the United States (the same geographical location as that of the majority of users in the MFP dataset). Interested workers had to take a qualification task by answering 5 survey questions designed to emulate the actual pairwise comparison tasks (see Sect. 4.2.2). The questions consist of a combination of identical food item pairs and similar food item pairs and workers were asked to assess their similarity by giving a rating from 1 (very dissimilar) to 5 (very similar). Workers who did not give a maximum rating to the identical pairs were disqualified. Likewise, workers who gave a higher rating to item pairs with fewer or no common textual features, e.g., (*milk*, *pizza*), than those with more common textual features, e.g., (*milk*, *chocolate milk*), were not qualified. At the end of this recruitment stage, 300 AMT workers successfully obtained the qualification to participate in the next stage.

Then, the qualified workers were instructed to recall and log food items they recently had in the past 3 days or longer using MyFitnessPal. The 3-day time window was specifically chosen based on the finding in the previous work (Liu et al. 2019) as the period likely to contain both repeat and novel consumptions. Moreover, they were required to log at least 3 food items per day. 50 AMT workers had fully complied with the instructions and were successfully enrolled into the online study as participants and allowed the research team to collect their food logging data. Later, data from 2



(a) Size of ground truth basket

(b) Number of recording days

Fig. 3 Data statistics from study II

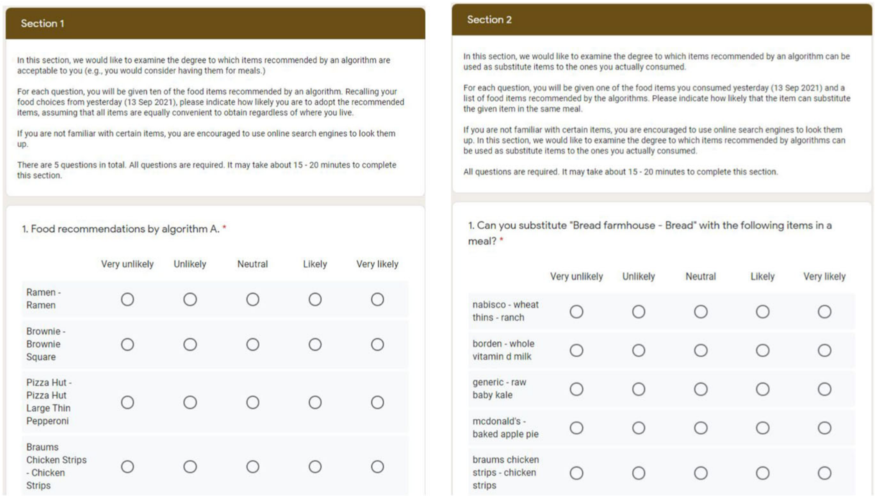
participants, who had completed the study, were removed since they were from the same MyFitnessPal account. Thus, we eventually used the data from 48 participants for the analysis.

Figure 3a displays a histogram of ground truth basket sizes from all participants. Most participants have 4–10 items in their ground truth baskets (mean = 7.6; S.D. = 4.32), whereas only two participants, who have been using MyFitnessPal actively prior to joining the study, have more than 20 items in their baskets. Figure 3b shows the distribution of the numbers of recording days of participants. The recording days may need not be continuous and we only count the days with food baskets recorded. Most participants have 3 to 4 recording days (mean = 6.1; S.D. = 10.13; median = 3.4). Two participants, who are active MyFitnessPal users, have logged food diaries for more than 50 days in the past year.

4.2.1 Part 1: next-basket recommendation

In this part of the survey, each participant was asked to rate their preference for each recommended item in the food item baskets. To generate the data for the study, we used the NBR models trained and evaluated in the offline experiment to recommend an item basket for the participants given their own food logging data. These users comprise an online test set (U_{online}). Similar to study I, we chose the same set of representative algorithms $ASet = \{$Personal, MixtureTW, NMF, FPMC, SASRec$\}$ to to generate a basket of top-10 recommended food items for each participant. Each model was trained and optimized following the procedures described in Sect. 3.4. For each user $u_i \in U_{online}$, his/her actual basket of food items logged on the day the participant joined the study (t) was used as the *ground truth basket* $G(i)$ and the remaining food logging data from days $t - 1$ onward were used as the test data. The top-10 recommended food items from all selected recommendation algorithms $ASet$ for user u_i is denoted as $RecList(i) = \bigcup_{algo \in ASet} RecList_{algo}(i, 10)$.

No later than 24 h after the participant had joined the study, an online survey (shown in Fig. 4a) was automatically created in Google Form and sent to the participant. The 24-hour limit was imposed to mimic the real-world user-system interactions manner and ensure that the participant's recall of his/her previous food choices was sufficiently reliable. In the survey, each participant was asked to rate how likely he/she was to adopt the recommended food items on a scale of 1 (very unlikely) to 5 (very likely). In total, the participant had to rate 50 items in $RecList_{algo}(i, 10)$, grouped into five 10-item



Section 1

In this section, we would like to examine the degree to which items recommended by an algorithm are acceptable to you (e.g., you would consider having them for meals.)

For each question, you will be given ten of the food items recommended by an algorithm. Recalling your food choices from yesterday (13 Sep 2021), please indicate how likely you are to adopt the recommended items, assuming that all items are equally convenient to obtain regardless of where you live.

If you are not familiar with certain items, you are encouraged to use online search engines to look them up.

There are 5 questions in total. All questions are required. It may take about 15 - 20 minutes to complete this section.

1. Food recommendations by algorithm A. *

	Very unlikely	Unlikely	Neutral	Likely	Very likely
Ramen - Ramen	○	○	○	○	○
Brownie - Brownie Square	○	○	○	○	○
Pizza Hut - Pizza Hut Large Thin Pepperoni	○	○	○	○	○
Braums Chicken Strips - Chicken Strips	○	○	○	○	○

(a) Next-basket recommendation survey

Section 2

In this section, we would like to examine the degree to which items recommended by an algorithm can be used as substitute items to the ones you actually consumed.

For each question, you will be given one of the food items you consumed yesterday (13 Sep 2021) and a list of food items recommended by the algorithms. Please indicate how likely that the item can substitute the given item in the same meal.

If you are not familiar with certain items, you are encouraged to use online search engines to look them up. In this section, we would like to examine the degree to which items recommended by algorithms can be used as substitute items to the ones you actually consumed.

All questions are required. It may take about 15 - 20 minutes to complete this section.

1. Can you substitute "Bread farmhouse - Bread" with the following items in a meal? *

	Very unlikely	Unlikely	Neutral	Likely	Very likely
nabisco - wheat thins - ranch	○	○	○	○	○
borden - whole vitamin d milk	○	○	○	○	○
generic - raw baby kale	○	○	○	○	○
mcdonald's - baked apple pie	○	○	○	○	○
braums chicken strips - chicken strips	○	○	○	○	○

(b) Basket-level similarity survey

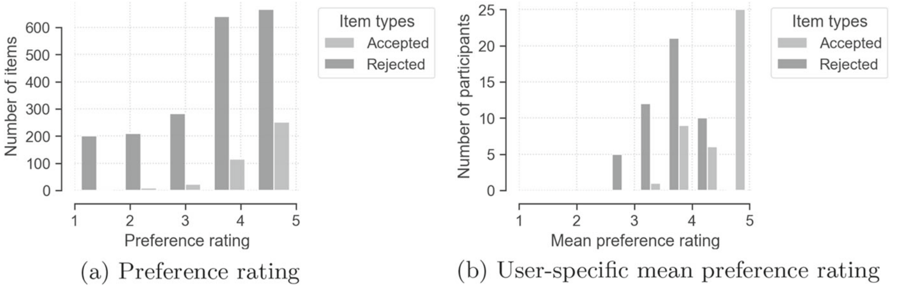
Fig. 4 Study II's surveys

baskets, i.e., one for each algorithm in $ASet$. The ordering of baskets and items in the survey was randomized to minimize the primacy effect. Moreover, algorithm names, used as the basket's titles, were also de-identified. 16.67% of all recommended items in the survey (400 of 2,400) were actually consumed items from the ground truth baskets, i.e., *accepted* items, whereas 83.33% of all recommended items (2,000 of 2,400) were those retroactively recommended but not actually consumed, i.e., *rejected* (Frumerman et al. 2019) or more precisely *non-accepted* items as the recommended items were actually presented to the participants shortly after the real food consumption decisions had been made. For convenience, the two terms, rejected and non-accepted, are used interchangeably in the discussion.

Let $r_i^p(v_j)$ denote a preference rating of user u_i for an item v_j and $\mu_r^p(i)$ denote a user-specific mean preference rating of u_i over the 50 recommended items rated by u_i in the survey, defined as:

$$\mu_r^p(i) = \frac{1}{50} \sum_{1 \le k \le 50} r_i^p(v_k)$$

Figure 5a and 5b displays the distributions of preference ratings and user-specific mean preference ratings from all participants, respectively. A vast majority of accepted items were given a rating of 4 or higher (92% of all preference ratings and 86% of all mean preference ratings), suggesting that the participants were quite attentive when answering the survey questions. Interestingly, nearly two third (65%) of all rejected items received a preference rating of 4 or higher. Similarly, 63% of all participants tend to lean slightly toward adopting some rejected items as indicated by their user-specific mean preference ratings of 3.0 to 4.0. The rating distributions suggest that



(a) Preference rating

(b) User-specific mean preference rating

Fig. 5 Distributions of preference ratings in the next-basket recommendation survey

some rejected items should not be treated as a complete failure when evaluating the effectiveness of the recommendations.

4.2.2 Part 2: personalized basket-level similarity survey

In the second part of the study, each participant u_i was instructed to assess the similarity between item pairs randomly sampled from a pool of $|G(i) \times RecList(i)|$ item pairs from the items he/she personally consumed ($G(i)$) and the ones recommended by the algorithms ($RecList(i)$). To mitigate the information overload problem (Malhotra 1982), each participant was assigned on average approximately 100 item pairs to judge. Note that five participants, who had taken part in an initial trial run, received a much larger number of item pairs (145–370 pairs) than the other participants. Given the item pairs data, we generated $|G(i)|$ survey questions, each corresponding to a ground truth item. In each question, the participant was asked to rate how likely each of the recommended items can be used as a substituted item to the ground truth item from 1 (very unlikely) to 5 (very likely). An example of basket-level similarity survey questions is shown in Fig. 4b. Together with the next-basket recommendation survey (part 1), the basket-level similarity survey was automatically generated in Google Form and sent to the participant within the first 24 h after he/she had joined the study. In total, the participants rated 5,458 item pairs.

Let $r_i^s(v_j, v_{j'})$ denote a substitution rating of user u_i for an item pair $(v_j, v_{j'})$, S_i denote a set of all item pairs rated by u_i, and $\mu_r^s(i)$ denote a user-specific mean substitution rating of u_i over all item pairs in S_i, defined as:

$$\mu_r^s(i) = \frac{1}{|S_i|} \sum_{1 \le k \le |S_i|} r_i^s(v_k, v_{k'})$$

Figure 6a and 6b displays the distributions of item-pair substitution ratings and user-specific mean substitution ratings from all participants, respectively. As we can see, 2,516 item pairs (46.1%) were given a rating of 1 (very unlikely substitutes). Of all item pairs, 75 pairs (1.37%) contain identical items, 92% of which were given a rating of 4 or higher (69.33% having a rating of 5). This indicates that the participants were reasonably attentive when performing the tasks. Next, most participants rated their

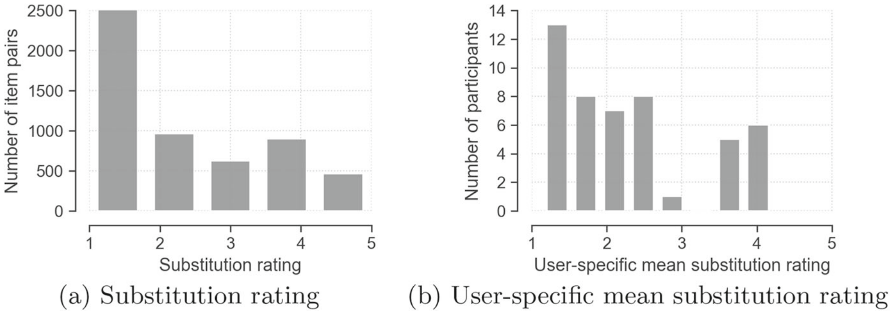


Fig. 6 Distributions of item pair substitution ratings in the basket-level similarity survey

item pairs with an average rating below 3.0, whereas 11 of 48 (22.92%) participants rated their item pairs with an average rating above 3.0.

5 Results and discussion

We present the analysis of the research data collected from the two user studies and answer the main research questions in this section. For convenience, we abbreviate the notations of all similarity measures and top-k evaluation metrics in this section, e.g., BLEU-1$(j'|j)$ and BLEU-1@10 are abbreviated as BLEU-1 in the respective contexts. To compute scores for all the content-based and hybrid metrics, we applied basic preprocessing steps to the food item data, including converting text into lower case and removing punctuation.

5.1 RQ1: How do different similarity metrics correspond to human similarity perception of items?

Using the pairwise similarity judgments data from the study I, we computed the Pearson's correlation coefficients (denoted by ρ) between the human similarity judgments (Sim_{human}) and the content-based and hierarchical item similarity scores for all 5,982 pairs. Furthermore, we included a baseline *identical* function that assigns $sim_{identical}(v_{j'}|v_j) = 1$ if v_j and $v_{j'}$ are identical; otherwise $sim_{identical}(v_{j'}|v_j) = 0$. The results are shown in Fig. 7a.

Overall, most hierarchical matching similarity metrics (hMatch-λ) correlate more strongly with Sim_{human} than the other metrics. Specifically, the two best metrics are hMatch-1 and hMatch-2; ρ(hMatch-1) = 0.4908 and ρ(hMatch-2) = 0.4795. This suggests that the AMT workers may partially rely on some form of implicit structured item semantics, as operationalized in hMatch-λ, when judging the similarity of item pairs more than relying on the textual content cues alone. While the hybrid hierarchical metrics hMatch$_{sim}$-λ variants outperform most content-based and embedding-based metrics on the correlations with human judgments, they did not outperform the best hierarchical matching metrics, for example, ρ(hMatch$_{sim}$-1) =

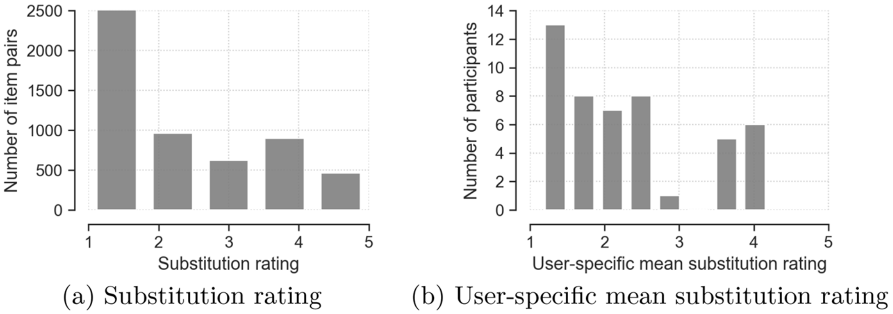


Fig. 7 Pearson's correlation scores between human judgments and different item similarity scores. All correlation scores are statistically significant ($p < 0.01$)

0.4218 and ρ(hMatch$_{sim}$-2) = 0.4184. Therefore, utilizing BERTscore for the tag-level similarity component in hMatch$_{sim}$-λ adversely affects its overall performance.

Among all content-based metrics, the n-gram-based similarity metrics generally correlate more strongly with Sim_{human} than the embedding-based similarity metrics, BLEU-1 being the metric with the highest correlation score (ρ(BLEU-1) = 0.4249) in this group and the 3rd best similarity metric overall. Surprisingly, all three BERTScore variants P$_{BERT}$, R$_{BERT}$, and F1$_{BERT}$ perform poorly in this task even though it has been shown that they outperform several similarity metrics, including n-gram-based metrics, in many NLP tasks (Zhang et al. 2020). Their correlation coefficients are slightly higher than that of the baseline identical metric; ρ(P$_{BERT}$) = 0.3386, ρ(R$_{BERT}$) = 0.3434, ρ(F1$_{BERT}$) = 0.3755, and ρ(identical) = 0.3104. Upon further inspection, we found that BERTScore (with a pretrained *RoBERTa large* model) tends to perform poorly given short food texts as inputs. For example, F1$_{BERT}$ scores for 3 following item pairs (*almond bars*, *cheese burger*), (*almond bars*, *roasted walnuts*), and (*almond bars*, *hot coffee*) are 0.2945, 0.2798, and 0.3133, respectively. However, one would intuitively expect (*almond bars*, *roasted walnuts*) to have the highest score among the three pairs instead of the lowest. This issue also likely explains the performances of the hMatch$_{sim}$-λ variants. Using BERTScore with a domain-specific BERT model fine-tuned on a food-related dataset may help improve the performance though we leave this to future work.

Next, results from the basket-level similarity survey in the study II, shown in Fig. 7b, are more or less consistent with those of the study I. That is: (1) hMatch-1 and hMatch-2 continue to correspond more closely to human judgments than the other metrics; (2) most hMatch$_{sim}$-λ variants are worse than hMatch-λ; and (3) most BERTScore variants are worse than most n-gram-based metrics. Nevertheless, the correlation gap of the best metric and the baseline identical metric is much smaller. Specifically, the correlation score of hMatch-1 is 35.88% higher than that of identical (ρ(hMatch-1) = 0.2727 and ρ(identical) = 0.2007) in the study II, compared to 58.09% in the study I (ρ(hMatch-1) = 0.4908 and ρ(identical) = 0.3104).

Since one of the main differences between the two studies is in the personalization of similarity judgments, we surmise that it partly contributes to the differences in the magnitude of the correlation coefficients. In particular, while the crowdsourced workers in the study I annotated each food item pair independent of the meal context and personal preference (i.e., non-personalized similarity judgments), the participants in the personalized basket-level similarity survey in the study II were expected to rely fairly on their personal preference and subjectivity when judging the item pairs given their own meal context (i.e., personalized similarity judgments). Thus, their similarity judgment ratings may be less homogeneous than those of the study I.

Summary. The proposed hierarchical matching functions hMatch-λ most correspond to human perception of item similarity, compared to the other similarity metrics. The magnitude of correlations significantly decreases as the human judgments become more personalized.

5.2 RQ2: How do different evaluation metrics correspond to the real users' preferences for item baskets?

With the preference ratings of 2,400 recommended items in 240 recommended baskets (48 participants × 5 recommended baskets by the five algorithms in $ASet$) collected in the next-basket recommendation survey, we examine the correlation coefficients (denoted by ρ) between the basket-level preference ratings and the evaluation scores computed for the recommended baskets. Let r_j^p be a preference rating of a recommended item v_j in a basket l. For each recommended basket l, we calculated its basket preference rating (denoted by $\mu_r^p(l)$) from an arithmetic mean of preference ratings r_j^p of the items belonging to the basket as follows:

$$\mu_r^p(l) = \frac{1}{|l|} \sum_{1 \leq j \leq |l|} r_j^p$$

We also computed evaluation scores using the standard, n-gram-based, embedding-based, hierarchical, and hybrid metrics for the recommended basket with respect to the corresponding ground truth basket. Finally, we computed Pearson's correlation coefficients for all pairs of (*basket preference ratings*, *evaluation scores*).

The results are shown in Fig. 8. As we can see, all content-based metrics, i.e., BLEU-N, ROUGE-N, and BERTScore, consistently outperform the other metrics, correlating most strongly with human preferences. Specifically, all three BERTScore variants have higher correlation with basket preference ratings than the other metrics; $\rho(\mathrm{P}_{BERT})$ = 0.5129, $\rho(\mathrm{R}_{BERT})$ = 0.4632, and $\rho(\mathrm{F1}_{BERT})$ = 0.5019. Surprisingly, precision, which has the highest correlation among all standard metrics, corresponds fairly well to the human preference ratings, ρ(precision) = 0.4438, even though it simply relies on the exact matching comparison between ground truth and recommended items. In contrast, all hierarchical and hybrid metrics do not correspond to human preference judgments better than the standard metrics even though their underlying hMatch functions are shown in RQ1's findings to correlate the most strongly with human similarity judgments. Among them, the recall-based metrics, e.g., hR-1, hR$_{sim}$-1, etc., greatly outperform the precision-based metrics, e.g., hP-1, hP$_{sim}$-1, etc. All hP$_{sim}$-λ variants have the lowest correlation with the human preference judgments; ρ(hP$_{sim}$-1) = 0.1604, ρ(hP$_{sim}$-2) = 0.1575, and ρ(hP$_{sim}$-IDF) = 0.1497.

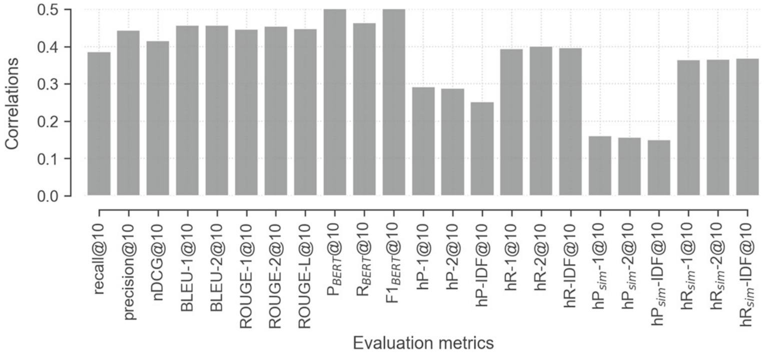


Fig. 8 Pearson's correlation scores between basket preference ratings and different top-k evaluation metrics. All correlation scores are statistically significant ($p < 0.05$)

Table 5 Accuracy of the selected top-k metrics for each preference rating quartile (Q). Best results are in bold face

Q	Precision	BLEU-2	P_{BERT}	hR-2	hR_{sim}-IDF
4 (most preferred baskets)	0.750	**0.754**	0.750	0.700	0.708
3	0.688	**0.717**	0.642	0.658	0.633
2	**0.671**	0.588	0.625	0.642	0.650
1 (least preferred baskets)	0.658	0.708	**0.750**	0.708	0.725

To better understand these results, we further examine the behaviors of selected evaluation metrics, especially in their effectiveness in distinguishing between highly preferred and less preferred baskets. That is, an ideal evaluation metric should assign proportionately high scores to highly preferred baskets, i.e, those with high preference ratings, and proportionately low scores to less preferred baskets, i.e., those with low preference ratings, most of the time. To characterize the performance metrics in this manner, we conducted a classification-based analysis. First, we chose five top-performing metrics from each group, i.e., precision, BLEU-2, P_{BERT}, hR-2, and hR_{sim}-IDF for comparison. Then, we split the preference ratings and evaluation scores into their respective quartiles for all 240 baskets. Baskets whose preference scores are in the top-25% (Q4; $\mu_r^p(l) = 5$; N = 56) are considered *most preferred* baskets, whereas those whose preference ratings are in the bottom-25% (Q1; $\mu_r^p(l) \leq 3.3$; N = 62) are considered *least preferred* baskets. Next, for each of the selected metrics, we constructed a confusion matrix for multi-class classification where the actual and predicted classes comprise the quartiles of the preference ratings and evaluation scores, respectively. A true positive (TP) case is met if the quartile of the preference rating of an item basket is the same as the quartile of the corresponding evaluation score. Similar logic is applied to false positive (FP), false negative (FN) and true negative (TN) cases. Then, from the confusion matrix, we computed accuracy ((TP+TN)/(TP+TN+FP+FN)), false positive rate (FP/(FP+TN)), and false negative rate (FN/(FN+TP)) for all preference rating quartiles.

Table 6 False positive and false negative rates of the selected top-k metrics for each preference rating quartile (Q). Best results are in bold face

Q	False Positive Rate					False Negative Rate				
	Precision	BLEU-2	P_{BERT}	hR-2	hR_{sim}-IDF	Precision	BLEU-2	P_{BERT}	hR-2	hR_{sim}-IDF
4	**0.120**	0.130	0.174	0.207	0.201	0.679	0.625	**0.500**	0.607	0.589
3	0.207	**0.168**	0.245	0.234	0.250	**0.661**	**0.661**	0.732	0.696	0.750
2	**0.144**	0.339	0.241	0.230	0.224	0.818	**0.606**	0.727	0.697	0.682
1	0.354	0.191	**0.163**	0.191	0.180	**0.306**	0.581	0.500	0.581	0.548

The results are displayed in Tables 5 and 6. Firstly, according to the accuracy scores, Precision, BLEU-2, and P_{BERT} are equally effective at assessing most preferred (Q4) baskets, whereas hR-2 and hR_{sim}-IDF are relatively less effective than the other metrics, notably due to having much higher false positive rates than the other three metrics. Secondly, all non-binary-based metrics are more accurate than the Precision metric in measuring least preferred (Q1) baskets. In particular, the false positive rate of Precision in measuring Q1 baskets is 0.354, 85.29% higher than that of P_{BERT}. Lastly, the accuracy of all metrics decreases when quantifying mid-range (Q2 and Q3) baskets. Within these groups of baskets, Precision and BLEU-2 tend to be more effective than P_{BERT}, hR-2, and hR_{sim}-IDF.

Summary. Most non-binary-based top-k metrics, especially the n-gram and embedding-based metrics, correspond more closely to human preference judgments than the standard metrics. Among those, P_{BERT} attains the highest correlation coefficient ($\rho(\text{P}_{BERT}) = 0.5129$). Most hierarchical and hybrid metrics correlate more poorly with human preference judgments than the other non-binary metrics and some standard metrics, i.e., precision and nDCG, despite the fact that their underlying similarity functions hMatch-λ correlate the strongest with human similarity judgments. Particularly, combining hierarchical matching with BERTScore adversely affects the discriminative power of both metrics against the user preference judgments. The personalized and preferential nature of the recommendation tasks may explain the differences. Lastly, all top-performing metrics are equally accurate in measuring highly preferred baskets. However, the precision metric is less accurate in measuring least preferred baskets than the non-binary-based counterparts.

5.3 RQ3: To what extent do user preferences for item baskets differ across different NBR algorithms?

In the next-basket recommendation survey, we collect from each participant a preference rating (from 1 to 5) for each of the top-10 recommended items by each of the five algorithms in $ASet$ = {Personal, MixtureTW, NMF, FPMC, SASRec}. Given an algorithm $algo$, we use $r_i^p(v_j)$ to denote the preference rating from user u_i on each item v_j in $RecList_{algo}(i, 10)$. We then derive the preference rating of user u_i on algorithm $algo$ by

$$r_i^p(algo) = \frac{1}{10} \sum_{v_j \in RecList_{algo}(i,10)} r_i^p(v_j)$$

By ordering the algorithms in $ASet$ by user preference ratings, we obtain the algorithm rank $rank^p(u_i, algo)$. Formally, $rank^p(u_i, algo)$ is defined as $rank^p(u_i, algo) = |\{algo' : r_i^p(algo') \geq r_i^p(algo), algo' \in ASet\}|$. When u_i gives the highest preference ratings to the algorithm $algo$, $rank^p(u_i, algo) = 1$. Then, the **mean preference ranking** of algorithm $algo$, denoted by $rank^p(algo)$, is defined by

$$rank^p(algo) = \frac{1}{|U|} \sum_{u_i \in U} rank^p(u_i, algo).$$

Table 7 Mean preference rankings of the selected algorithms. Lower is better

Personal	MixtureTW	NMF	FPMC	SASRec
1.90	**1.44**	3.52	3.88	3.44

As shown in Table 7, the basket recommendations of MixtureTW achieve the best mean preference ranking (1.44), whereas those of FPMC receive the worst mean preference ranking (3.88). Next, the Kruskal–Wallis test suggests that the median of preference rankings of the five algorithms are statistically different (p <0.05). Specifically, the post hoc multiple comparison tests using Dunn's test with Bonferroni correction show six paired comparisons which are statistically different, i.e., (Personal, NMF), (Personal, FPMC), (Personal, SASRec), (MixtureTW, NMF), (MixtureTW, FPMC), and (MixtureTW, SASRec). There are no differences in the mean preference rankings between MixtureTW and Personal, or among NMF, FPMC, and SASRec.

As we can see in Table 1, more than 50% of food consumptions in the MFP dataset consist of repeat consumptions. The main contributor to the results is therefore due to the effectiveness of the two repeat consumption-aware algorithms (i.e., Personal and MixtureTW) in recommending items the participants were likely to consume again. Specifically, since the result of Personal does not differ statistically from that of MixtureTW, we could attribute the success of both methods virtually to the repeat items recommendation over the novel items recommendation. In contrast, the more sophisticated sequence-aware algorithms (i.e., FPMC and SASRec), which learn to predict next baskets from the past sequences of baskets, were not able to effectively capture the dynamics of repeat-novel consumptions in the basket data.

Summary Among the five representative algorithms, most participants prefer item baskets recommended by MixtureTW and Personal over SASRec, NMF, and FPMC. The findings call to attention the challenge of the next-basket food recommendation task in which item baskets recommended by relatively simpler methods, such as Personal and MixtureTW, are generally more preferred by real users than those recommended by more sophisticated methods, such as FPMC and SASRec.

5.4 RQ4: What is the offline performance of different NBR algorithms as measured by the non-binary evaluation metrics?

Using the same trained models, predicted baskets, and test set described in Sect. 3.4, we computed performance scores using the selected representative evaluation metrics which were found in RQ2 to strongly correlate with human preference judgments, i.e., precision, BLEU-2, P$_{BERT}$, hR-2, and hR$_{sim}$-IDF, for all 12 recommender algorithms. The offline evaluation results are shown in Table 8. Firstly, MixtureTW performs the best among all the algorithms across all metrics, whereas the random baseline performs the worst in all metrics except for hR$_{sim}$-IDF. Secondly, the latent factor-based algorithms consistently perform at the bottom-50% across all metrics. Thirdly, the relative performances of certain algorithms are judged differently by precision compared to those by P$_{BERT}$. For example, the Personal baseline under-performs by 2 ranks, whereas FPMC and BPR-MF over-perform by 2 ranks when comparing the

Table 8 Offline performance of different algorithms with the selected metrics on the hold-out MFP test set. Best results are in bold face

Method	Scores ↑					Ranks ↓				
	Precision	BLEU-2	P_{BERT}	hR-2	hR_{sim}-IDF	Precision	BLEU-2	P_{BERT}	hR-2	hR_{sim}-IDF
Random	0	0.003	0.141	0.261	0.668	12	12	12	12	9
Global	0.031	0.03	0.214	0.276	0.619	10	11	11	11	12
Personal	0.134	0.112	0.304	0.585	0.685	4	5	2	2	7
Mixture	0.135	0.142	0.290	0.532	0.774	3	2	3	3	2
MixtureTW	**0.165**	**0.169**	**0.310**	**0.594**	**0.813**	**1**	**1**	**1**	**1**	**1**
adaLoyal	0.127	0.111	0.267	0.492	0.750	5	6	6	5	4
NMF	0.061	0.083	0.248	0.390	0.684	8	7	7	7	8
BPR-MF	0.062	0.06	0.244	0.366	0.689	7	8	9	8	6
WRMF	0.054	0.055	0.244	0.333	0.658	9	9	8	9	10
LDA	0.031	0.034	0.216	0.325	0.640	10	10	10	10	11
FPMC	0.143	0.129	0.286	0.513	0.766	2	3	4	4	3
SASRec	0.113	0.12	0.275	0.481	0.743	6	4	5	6	5

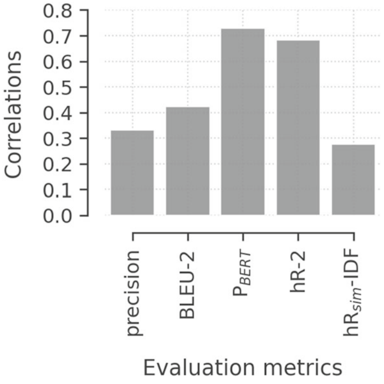


Fig. 9 Spearman's correlation scores between human preferences and top-*k* evaluation metrics

rankings from precision versus P_{BERT}. Interestingly, hR_{sim}-IDF is the only metric that greatly overestimates the performance of the Random baseline, which highlights its drawback as a reliable evaluation metric.

Next, using the preference rankings of the 5 representative algorithms from the user study II as ground truth (i.e., rank(MixtureTW) = 1, rank(Personal) = 2, rank(SASRec) = 3, rank(NMF) = 4, and rank(FPMC)=5), we computed Spearman's rank correlation scores (ρ_s) between the ground truth rankings and the rankings from the corresponding metrics in Table 8. As shown in Fig. 9, the rankings from P_{BERT} and hR-2 correspond more closely to the human preferences ranking than those from other metrics; $\rho_s(\text{P}_{BERT})$=0.7285, ρ_s(hR-2)=0.6822, ρ_s(BLEU-2)=0.4243, ρ_s(precision)=0.3313, and $\rho_s(\text{hR}_{sim}$-IDF)=0.2761. Note that due to a very small sample size (the number of NBR algorithms being ranked), ρ_s are not statistically significant ($p > 0.05$). Nevertheless, the correlation scores still provide a useful and reliable measurement of the similarity between the preference rankings (Fagin et al. 2003).

Summary: MixtureTW is the top performing algorithm across all metrics in the offline experiment. When comparing against human judgments of selected algorithms, P_{BERT} and hR-2 produce the performance rankings that correspond most closely to the ground truth ranking obtained from the study participants in the online user study. Overall, the offline performance assessment from the non-binary metrics is more consistent with the online experiment performance than that of the standard binary-based metrics.

6 Limitations and future directions

We acknowledge a few limitations in our research and discuss potential directions for future research in the followings.

6.1 Generalizability to other NBR domains

Although this study focuses solely on the next-meal recommendation in the food consumption domain, we believe the results of the offline evaluation are generally applicable to other NBR domains, such as grocery shopping, music listening, etc., due to the characteristics of the consumption data commonly shared among them. That is, the users in those domains tend to consume a collection of repeat and novel items in a basket. According to the offline evaluation results conducted on three other NBR datasets in appendix A, the repeat consumption-aware methods are the most effective in the NBR tasks across all datasets and evaluation metrics. Furthermore, relatively simple algorithms, such as Personal and MixtureTW, outperform more sophisticated/deep-learning-based methods, such as adaLoyal, FPMC, and SASRec. As the proportions of repeat consumption in the datasets are high, the overall NBR performance tends to be dominated by the methods which perform better on the repeat consumption task. The algorithmic performance may differ in other NBR domains where repeat consumption is not as prevalent as the food consumption, grocery shopping, and music listening domains.

6.2 Applicability of non-binary metrics

The non-binary-based metrics proposed in this work, including the content-based and hierarchical evaluation metrics, require the item data with textual contents and/or categorical descriptions to quantitatively assess the quality of the recommended baskets. In some NBR datasets where such information is not available, it is not possible to perform the non-binary relevance assessment of the recommendations. For demonstration, we have provided additional offline evaluation in which the non-binary evaluation metrics were utilized with other NBR datasets in appendix A.

6.3 BERTScore and food-domain knowledge

The effectiveness of BERTScore in several NLP evaluation tasks is due in large part to a massive amount of language and world knowledge contained in the pre-trained BERT models. Nevertheless, the results from our item similarity user study show that using a default *RoBERTa* model (Liu et al. 2019) with BERTScore is not an optimal setup. To improve the discriminative power of BERTScore in evaluating the food item similarity task, domain adaptation techniques (Rietzler et al. 2020; Ma et al. 2019) can be used to better adapt the pre-trained BERT model. For instance, the pre-trained weights of the default BERT model could further be fine-tuned on a large food and recipe-specific corpus like Recipe1M+ (Marin et al. 2019) in a self-supervised manner (i.e., pre-training). The fine-tuned food-domain BERT model could then replace the default *RoBERTa* in BERTScore. Other knowledge infusion techniques for pre-trained language models (He et al. 2020; Penha and Hauff 2020) could be explored to enrich or inject BERT with the food-domain knowledge.

A few issues should also be taken into consideration when fine-tuning BERT. Firstly, the fine-tuned BERT models have been shown to produce inconsistent performance due to the instability during the fine-tuning process (Mosbach et al. 2020; Zhang et al. 2020) and more systematic studies (Ganesan et al. 2021) are still needed to investigate the causes of the fine-tuning instability. Future research should carefully consider the issue and empirically evaluate how different fine-tuning methods affect

Table 9 Dataset statistics

	#users	#items	#transactions	density	#baskets	#items per user	basket size	%repeat consumption
Dunnhumby	2,148	12,763	2,115,821	3.84%	208,320	490.22 ± 311.44	10.16 ± 11.95	41.07% ± 15.75%
Instacart	206,209	49,685	33,819,106	0.14%	3,346,083	67.23 ± 56.88	10.11 ± 7.54	45.05% ± 20.21%
LastFM	10,371	39,203	94,147,706	1.19%	6,372,465	467.68 ± 424.25	5.26 ± 8.81	79.79% ± 14.81%
MFP	6,916	47,789	2,260,319	0.23%	414,874	107.68 ± 79.8	5.45 ± 3.39	55.69% ± 18.77%

the performance of the fine-tuned BERT models in the food item similarity task. Next, an optimally fine-tuned BERTScore metric could potentially achieve a human-level performance in the item similarity task; however, the state-of-the-art performance likely comes at the expense of a significant increase in computational cost. In some cases, the trade-off between performance and computational cost may result in up to 100 times more training time to optimize a fine-tuned BERT-based metric than conventional non-BERT-based metrics (Mayfield and Black 2020). Therefore, the opportunity cost of fine-tuning BERT should be taken into account as a matter of practicality, especially in an environment with limited computing resources.

6.4 Assessment of basket qualities

All non-binary-based evaluation metrics proposed and investigated in this research are operationalized based on the *best matching principle* in which a basket-level performance score is aggregated from multiple comparisons between individual items in the recommended basket and those in the ground truth basket. There are a few advantages for this approach. Firstly, the metrics are computationally efficient due to the aggregation operation. Next, they generally follow the offline evaluation paradigm, in which the recommendations are compared against the ground truths, and are therefore easy to understand. Moreover, since they do not incorporate any domain-specific qualities into the assessment, they can be applied to other NBR domains.

On the other hand, our proposed non-binary-based metrics, which are accuracy-oriented, are not designed to measure other domain-specific characteristics of the recommendations in the food consumption and related NBR domains (e.g., grocery shopping). Specifically, since baskets/meals typically comprise food items which are meant to be consumed together, one could reasonably assume that users are likely to prefer recommended baskets with more complementary items to those with fewer complementary items. For example, a user who prefers balanced diet may be more satisfied with a meal recommendation of $\{steak, pasta, salad\}$ (more complementary) than a recommendation of $\{steak, pork\ chop, chicken\ wings\}$ (less complementary). Following the non-accuracy metrics in recommender systems research (Ge et al. 2010), future work should incorporate the within-basket item relationships, such as complementarity and substitutability (Achananuparp and Weber 2016), in the NBR evaluation beyond accuracy.

6.5 User studies

The environments of our user studies may differ from the environments of live recommender systems where users are free to interact with the recommendations. Even though we have tried to mimic the production recommender systems in the user study, i.e., by generating and presenting personalized basket recommendations to the participant as soon as he/she has submitted their past meal history, user perceptions of the recommendations in a controlled environment may still not necessarily be the same as those in a live environment.

Our online user study was conducted in the next-meal recommendation. To further examine how the non-binary evaluation paradigm can generalize to other application domains, future research could consider conducting online evaluations and user studies in related NBR domains, such as grocery shopping. Lastly, stronger evidence from future larger-scale longitudinal user studies (Achananuparp et al. 2018; Hauptmann et al. 2021) that captures the temporal food consumption patterns (e.g., weekdays vs. weekends) (Liu et al. 2019) could help further validate our research findings.

6.6 Quality of research data

Lastly, we collected the food consumption data and survey responses from qualified crowdworkers. Therefore, the data quality in the recommendation research may vary (Musto et al. 2020; Trattner and Jannach 2020). Nevertheless, we believe that our workers selection criteria, qualification test, attention check, and data verification have sufficiently helped limit the validity risk.

7 Conclusion

This research aims to broaden knowledge on the evaluation of next-basket recommendation (NBR) in the food recommendation domain. In particular, we investigated the non-binary relevance assessment in measuring the quality of recommended item baskets based on our claim that partial credits should be given to the recommended baskets which share some similarity to the ground truth. We proposed various non-binary-based metrics for item-level and basket-level measurements by adapting and extending relevant similarity metrics used in the natural language processing (NLP) and text classification research, including BLEU, ROUGE, BERTScore, and hierarchical evaluation metrics. Next, we validated the proposed non-binary-based metrics using a large food diary dataset and several state-of-the-art NBR algorithms in the online and offline experiments. Specifically, two user studies were conducted via the Amazon Mechanical Turk platform to obtain human judgments of basket similarity and preference.

We identified a few key findings from the experimental results. Firstly, among all non-binary-based item similarity metrics, the hierarchical matching function hMatch-λ correlates the most strongly with human judgments of item similarity. This indicates its potential in the non-binary NBR evaluation. Secondly, the majority of non-binary-based top-k metrics correlate more strongly with human preference judgments than the binary-based metrics. Among the non-binary-based metrics, an embedding-based metric P_{BERT} has the highest correlation with human preference judgments. Surprisingly, most hierarchical evaluation metrics have lower correlations with human preference judgments than the others even though they most strongly correlate with human judgments of item similarity. Next, results from the online next-meal recommendation user study show that the participants generally prefer the basket recommendations from the repeat-consumption aware methods (MixtureTW and Personal) over the recommendations from the more sophisticated sequential recommendation algorithms

(SASRec and FPMC). Lastly, according to the online and offline experiments, non-binary evaluation metrics, such as P_{BERT} and hR-2, are more indicative of the online experiment performance than precision, suggesting the validity of the non-binary relevance assessment and the limitations of standard binary-based metrics in the offline NBR evaluation.

Acknowledgements This research is supported by the National Research Foundation, Singapore, under its International Research Centres in Singapore Funding Initiative. Any opinions, findings and conclusions, or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

Funding This research is supported by the National Research Foundation, Singapore, under its International Research Centres in Singapore Funding Initiative.

Availability of data and materials Not applicable.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

Ethical approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Code availability Not applicable.

Appendix A: Results for other NBR domains

We conducted additional offline evaluation using three other grocery shopping and music listening datasets commonly used in other NBR domains: Dunnhumby,[1] Instacart,[2] and LastFM.[3]

The **Dunnhumby** dataset contains offline household-level grocery transactions over two years collected from around 2K households. The users are frequent shoppers with an average shopping frequency of once per week. Each item has a unique ID and is associated with one of the 27 departments (e.g., produce), one of the 249 descriptions (e.g., tropical fruit), and one of the 1,138 specific descriptions (e.g., bananas). In the p-core filtering, min_{user} and min_{item} are both 10.

The **Instacart** dataset was published by an online grocery service Instacart.com in the United States. It contains over 3 M grocery orders from more than 200K users. Although the specific date of each order is not provided in the dataset, the sequence of transactions by each user is preserved. Each item has a unique name and is associated with one of the 21 departments (e.g., produce) and one of the 134 aisles (e.g., fresh fruits). In the p-core filtering, min_{user} and min_{item} are both 20.

[1] https://www.dunnhumby.com/sourcefiles.

[2] https://www.instacart.com/datasets/grocery-shopping-2017.

[3] http://www.cp.jku.at/datasets/LFM-1b/.

The **LastFM** 1 Billion dataset (Schedl and Ferwerda 2017) contains 1B music listening events. We focus on the 219,589 artists annotated with one or more tags from the 20 genre tags (e.g., classical, electronic, etc.) from an online music portal AllMusic.com. We define a transaction as a user listening to an artist and a new basket for a user (i.e., listening session) is defined if the interval between consecutive songs is greater than an hour. In the p-core filtering, min_{user} and min_{item} are both 10.

For all three datasets, we applied the same data preprocessing, experimental setup, and protocols used in the main study as described in Sect. 3. After data preprocessing, their basic statistics are described in Table 9. Similar to the MFP dataset, these datasets are also highly sparse and contain a large degree of repeat consumption.

The performance scores were computed using nDCG@10 for the binary evaluation metric and hR2-@10 and hR-1@10 for the non-binary evaluation metrics. All three datasets contain relevant metadata such as category hierarchies and tags. However, none of them contains the items' full-text descriptions. Therefore, the only applicable non-binary evaluation metric for the experiment is hierarchical evaluation metrics. Specifically, we computed hR-2@10 for Dunnhumby and Instacart as both datasets have a full category hierarchy, whereas we computed hR-1@10 for LastFM due to a flat structure of tags. Both hR-2@10 and hR-1@10 have shown to correlate more strongly with human judgments, compared to the other variants.

The results are shown in Table 10. Here, the effectiveness of different algorithms NBR methods on the three other NBR datasets is generally similar to the results from the main study. Specifically, the repeat consumption-aware methods, such as Personal and MixtureTW, outperform the more sophisticated algorithms. In both the Dunnhumby and Instacart datasets, MixtureTW is the best in terms of nDCG@10, whereas Personal is the best in terms of hR-2@10. The performance difference between MixtureTW and Personal is mainly due to a poor novel item recommendation performance of MixtureTW, resulting in more irrelevant tags negatively affecting its hR-2@10 scores. Next, for the Dunnhumby dataset, most latent factor-based algorithms do not outperform the naive Global baseline in terms of nDCG@10. However, their hR-2@10 scores are all higher than those of Global. This suggests that although the items in the recommended baskets are not identical to the ones in the ground truth baskets, their baskets share more commonality in terms of attributes and categories.

Next, the hierarchical metric scores for the LastFM dataset are much higher (ranging from 0.734–0.94) than those of the other two datasets. This can be explained by the fact that the number of tags in LastFM is much smaller (only 20 music genres) than those of Dunnhumby and Instacart. Therefore, it is much more likely for the algorithms to recommend one relevant tag and receive higher hierarchical evaluation scores (hR-1@10) in the LastFM experiment. Even a naive Random baseline can recall over 73.4% of the tags.

Table 10 Offline performance of NBR methods on Dunnhumby, Instacart and LastFM datasets. Best results are in bold face

	Dunnhumby		Instacart		LastFM	
	nDCG@10	hR-2@10	nDCG@10	hR-2@10	nDCG@10	hR-1@10
Random	0.002	0.200	0.000	0.178	0.000	0.734
Global	0.109	0.263	0.059	0.198	0.016	0.764
Personal	0.208	**0.404**	0.110	**0.272**	0.192	0.867
Mixture	0.207	0.368	0.112	0.259	0.184	0.865
MixtureTW	**0.224**	0.386	**0.128**	0.262	**0.357**	**0.940**
adaLoyal	0.188	0.366	0.101	0.253	0.333	0.905
NMF	0.109	0.341	0.050	0.207	0.085	0.840
BPR-MF	0.093	0.283	0.091	0.226	0.059	0.821
WRMF	0.079	0.277	0.088	0.227	0.062	0.845
LDA	0.109	0.279	0.060	0.204	0.025	0.812
FPMC	0.144	0.296	0.092	0.237	0.061	0.825
SASRec	0.150	0.349	0.093	0.242	0.064	0.828

References

Achananuparp, P., Hu, X., Shen, X.: The evaluation of sentence similarity measures. In: International Conference on Data Warehousing and Knowledge Discovery, pp. 305–316 (2008). https://doi.org/10.1007/978-3-540-85836-2_29. Springer

Achananuparp, P., Hu, X., Yang, C.C.: Addressing the variability of natural language expression in sentence similarity with semantic structure of the sentences. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 548–555 (2009). https://doi.org/10.1007/978-3-642-01307-2_52. Springer

Achananuparp, P., Lim, E.-P., Abhishek, V., Yun, T.: Eat & tell: A randomized trial of random-loss incentive to increase dietary self-tracking compliance. In: Proceedings of the 2018 International Conference on Digital Health. DH'18, pp. 45–54. ACM, New York, NY, USA (2018). https://doi.org/10.1145/3194658.3194662

Achananuparp, P., Weber, I.: Extracting food substitutes from food diary via distributional similarity. CoRR (2016). arXiv:1607.08807

Arora, S., Liang, Y., Ma, T.: A simple but tough-to-beat baseline for sentence embeddings. In: International Conference on Learning Representations (2017)

Beel, J., Genzmehr, M., Langer, S., Nürnberger, A., Gipp, B.: A comparative analysis of offline and online evaluations and discussion of research paper recommender system evaluation. In: Proceedings of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation, pp. 7–14 (2013). https://doi.org/10.1145/2532508.2532511

Beel, J., Breitinger, C., Langer, S., Lommatzsch, A., Gipp, B.: Towards reproducibility in recommender-systems research. User Model. User-Adapted Interact. **26**(1), 69–101 (2016). https://doi.org/10.1007/s11257-016-9174-x

Bharadhwaj, H., Park, H., Lim, B.Y.: Recgan: Recurrent generative adversarial networks for recommendation systems. In: Proceedings of the 12th ACM Conference on Recommender Systems, pp. 372–376 (2018). https://doi.org/10.1145/3240323.3240383

Braunhofer, M., Kaminskas, M., Ricci, F.: Location-aware music recommendation. Int. J. Multimed. Inf. Retriev. **2**(1), 31–44 (2013). https://doi.org/10.1007/s13735-012-0032-2

Brilhante, I., Macedo, J.A., Nardini, F.M., Perego, R., Renso, C.: Where shall we go today? planning touristic tours with tripbuilder. In: Proceedings of the 22nd ACM International Conference on Information & Knowledge Management, pp. 757–762 (2013). https://doi.org/10.1145/2505515.2505643

Chen, S., Moore, J.L., Turnbull, D., Joachims, T.: Playlist prediction via metric embedding. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 714–722 (2012). https://doi.org/10.1145/2339530.2339643

Cheng, C., Yang, H., Lyu, M.R., King, I.: Where you like to go next: Successive point-of-interest recommendation. In: Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence, pp. 2605–2611 (2013)

Colucci, L., Doshi, P., Lee, K.-L., Liang, J., Lin, Y., Vashishtha, I., Zhang, J., Jude, A.: Evaluating item-item similarity algorithms for movies. In: Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems, pp. 2141–2147 (2016). https://doi.org/10.1145/2851581.2892362

Cremonesi, P., Garzotto, F., Turrin, R.: Investigating the persuasion potential of recommender systems from a quality perspective: an empirical study. ACM Trans. Interact. Intell. Syst. **2**(2), 1–41 (2012). https://doi.org/10.1145/2209310.2209314

Dacrema, M.F., Cremonesi, P., Jannach, D.: Are we really making much progress? a worrying analysis of recent neural recommendation approaches. In: Proceedings of the 13rd ACM Conference on Recommender Systems, pp. 101–109 (2019). https://doi.org/10.1145/3298689.3347058

Delice, A.: The sampling issues in quantitative research. Educ. Sci. Theory Pract. **10**(4), 2001–2018 (2010)

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

Domingues, M.A., Gouyon, F., Jorge, A.M., Leal, J.P., Vinagre, J., Lemos, L., Sordo, M.: Combining usage and content in an online recommendation system for music in the long tail. Int. J. Multimed. Inf. Retriev. **2**(1), 3–13 (2013)

Elsweiler, D., Hauptmann, H., Trattner, C.: In: Ricci, F., Rokach, L., Shapira, B. (eds.) Food Recommender Systems, pp. 871–925. Springer, New York, NY (2022). https://doi.org/10.1007/978-1-0716-2197-4_23

Faggioli, G., Polato, M., Aiolli, F.: Recency aware collaborative filtering for next basket recommendation. In: Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization, pp. 80–87 (2020). https://doi.org/10.1145/3340631.3394850

Fagin, R., Kumar, R., Sivakumar, D.: Comparing top k lists. SIAM J. Discrete Math. **17**(1), 134–160 (2003)

Färber, M., Sampath, A.: Hybridcite: A hybrid model for context-aware citation recommendation. In: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020, pp. 117–126 (2020). https://doi.org/10.1145/3383583.3398534

Frumerman, S., Shani, G., Shapira, B., Sar Shalom, O.: Are all rejected recommendations equally bad? Towards analysing rejected recommendations. In: Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization, pp. 157–165 (2019). https://doi.org/10.1145/3320435.3320448

Garcin, F., Faltings, B., Donatsch, O., Alazzawi, A., Bruttin, C., Huber, A.: Offline and online evaluation of news recommender systems at swissinfo. ch. In: Proceedings of the 8th ACM Conference on Recommender Systems, pp. 169–176 (2014). https://doi.org/10.1145/2645710.2645745

Ge, M., Delgado-Battenfeld, C., Jannach, D.: Beyond accuracy: evaluating recommender systems by coverage and serendipity. In: Proceedings of the fourth ACM Conference on Recommender Systems—RecSys'10, pp. 257 (2010). https://doi.org/10.1145/1864708.1864761

Ge, M., Elahi, M., Fernaández-Tobías, I., Ricci, F., Massimo, D.: Using tags and latent factors in a food recommender system. In: Proceedings of the 5th International Conference on Digital Health 2015, pp. 105–112 (2015). https://doi.org/10.1145/2750511.2750528

Gopalan, P., Hofman, J.M., Blei, D.M.: Scalable recommendation with hierarchical Poisson factorization. In: Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence, pp. 326–335 (2015)

Hauptmann, H., Leipold, N., Madenach, M., Wintergerst, M., Lurz, M., Groh, G., Böhm, M., Gedrich, K., Krcmar, H.: Effects and challenges of using a nutrition assistance system: results of a long-term mixed-method study. User Model. User-Adapted Interact. 1–53 (2021). https://doi.org/10.1007/s11257-021-09301-y

He, H., Gimpel, K., Lin, J.: Multi-perspective sentence similarity modeling with convolutional neural networks. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 1576–1586 (2015). https://doi.org/10.18653/v1/D15-1181

He, J., Li, X., Liao, L.: Category-aware next point-of-interest recommendation via listwise bayesian personalized ranking. In: IJCAI, vol. 17, pp. 1837–1843 (2017)

He, Y., Zhu, Z., Zhang, Y., Chen, Q., Caverlee, J.: Infusing disease knowledge into bert for health question answering, medical inference and disease name recognition. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 4604–4614 (2020)

Hidasi, B., Karatzoglou, A., Baltrunas, L., Tikk, D.: Session-based Recommendations with Recurrent Neural Networks. arXiv (2015). https://doi.org/10.48550/ARXIV.1511.06939

Hu, H., He, X., Gao, J., Zhang, Z.-L.: Modeling personalized item frequency information for next-basket recommendation. Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1071–1080 (2020). https://doi.org/10.1145/3397271.3401066

Hu, H., He, X.: Sets2sets: Learning from sequential sets with neural networks. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 1491–1499 (2019). https://doi.org/10.1145/3292500.3330979

Hu, Y., Koren, Y., Volinsky, C.: Collaborative filtering for implicit feedback datasets. In: Proceedings of the 8th IEEE International Conference on Data Mining, pp. 263–272 (2008). https://doi.org/10.1109/ICDM.2008.22

Huang, G., Guo, C., Kusner, M.J., Sun, Y., Sha, F., Weinberger, K.Q.: Supervised word mover's distance. Advances in neural information processing systems vol. 29 (2016)

Jannach, D., Ludewig, M.: When recurrent neural networks meet the neighborhood for session-based recommendation. Proceedings of the Eleventh ACM Conference on Recommender Systems, 306–310 (2017). https://doi.org/10.1145/3109859.3109872

Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of ir techniques. ACM Trans. Inf. Syst. **20**(4), 422–446 (2002). https://doi.org/10.1145/582415.582418

Jones, S.L., Kelly, R.: Dealing with information overload in multifaceted personal informatics systems. Hum. Comput. Interact. **33**(1), 1–48 (2018). https://doi.org/10.1080/07370024.2017.1302334

Kamehkhosh, I., Jannach, D.: User perception of next-track music recommendations. In: Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization, pp. 113–121 (2017). https://doi.org/10.1145/3079628.3079668

Kaminskas, M., Bridge, D., Foping, F., Roche, D.: Product recommendation for small-scale retailers. In: International Conference on Electronic Commerce and Web Technologies, pp. 17–29 (2015). https://doi.org/10.1007/978-3-319-27729-5_2. Springer

Kang, W.-C., McAuley, J.: Self-attentive sequential recommendation. In: 2018 IEEE International Conference on Data Mining (ICDM), pp. 197–206 (2018). https://doi.org/10.1109/ICDM.2018.00035

Kapoor, K., Subbian, K., Srivastava, J., Schrater, P.: Just in time recommendations: modeling the dynamics of boredom in activity streams. In: Proceedings of the Eighth ACM International Conference on Web Search Data Min. - WSDM'15, pp. 233–242 (2015). https://doi.org/10.1145/2684822.2685306

Kenter, T., De Rijke, M.: Short text similarity with word embeddings. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, pp. 1411–1420 (2015). https://doi.org/10.1145/2806416.2806475

Kiritchenko, S., Matwin, S., Famili, A.F., et al.: Functional annotation of genes using hierarchical text categorization. In: Proceedings of the ACL Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics (2005)

Kiros, R., Zhu, Y., Salakhutdinov, R.R., Zemel, R., Urtasun, R., Torralba, A., Fidler, S.: Skip-thought vectors. Advances in neural information processing systems, vol. 28 (2015)

Kotzias, D., Lichman, M., Smyth, P.: Predicting consumption patterns with repeated and novel events. IEEE Trans. Knowl. Data Eng. **31**(2), 371–384 (2019). https://doi.org/10.1109/TKDE.2018.2832132

Krauth, K., Dean, S., Zhao, A., Guo, W., Curmei, M., Recht, B., Jordan, M.I.: Do offline metrics predict online performance in recommender systems? arXiv preprint arXiv:2011.07931 (2020)

Kusner, M., Sun, Y., Kolkin, N., Weinberger, K.: From word embeddings to document distances. In: International Conference on Machine Learning, pp. 957–966 (2015). PMLR

Lacic, E., Kowald, D., Theiler, D., Traub, M., Kuffer, L., Lindstaedt, S.N., Lex, E.: Evaluating tag recommendations for e-book annotation using a semantic similarity metric. CoRR (2019). arXiv:1908.04042

Le, D.-T., Lauw, H.W., Fang, Y.: Correlation-sensitive next-basket recommendation. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19, pp. 2808–2814 (2019). https://doi.org/10.24963/ijcai.2019/389

Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: International Conference on Machine Learning, pp. 1188–1196 (2014). PMLR

Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: Proceedings of the 13th International Conference on Neural Information Processing Systems, pp. 535–541 (2000)

Li, M., Jullien, S., Ariannezhad, M., de Rijke, M.: A Next Basket Recommendation Reality Check. arXiv e-prints (2021) arXiv:2109.14233 [cs.IR]

Li, Y., McLean, D., Bandar, Z.A., O'shea, J.D., Crockett, K.: Sentence similarity based on semantic nets and corpus statistics. IEEE Trans. Knowl. Data Eng. **18**(8), 1138–1150 (2006). https://doi.org/10.1109/TKDE.2006.130

Lin, C.-Y.: ROUGE: A package for automatic evaluation of summaries. Text Summarization Branches Out, 74–81 (2004)

Liu, Y., Lee, H., Achananuparp, P., Lim, E.-P., Cheng, T.-L., Lin, S.-D.: Characterizing and predicting repeat food consumption behavior for just-in-time interventions. In: Proceedings of the 9th International Conference on Digital Public Health, pp. 11–20 (2019). https://doi.org/10.1145/3357729.3357736

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv (2019). https://doi.org/10.48550/ARXIV.1907.11692

Lops, P., Gemmis, M.d., Semeraro, G.: Content-based recommender systems: State of the art and trends. Recommender Systems Handbook, pp. 73–105 (2011). https://doi.org/10.1007/978-0-387-85820-3_3

Ma, X., Xu, P., Wang, Z., Nallapati, R., Xiang, B.: Domain adaptation with BERT-based domain classification and data selection. In: Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019), pp. 76–83. Association for Computational Linguistics, Hong Kong, China (2019). https://doi.org/10.18653/v1/D19-6109

Maksai, A., Garcin, F., Faltings, B.: Predicting online performance of news recommender systems through richer evaluation metrics. In: Proceedings of the 9th ACM Conference on Recommender Systems, pp. 179–186 (2015). https://doi.org/10.1145/2792838.2800184

Malhotra, N.K.: Information load and consumer decision making. J. Consum. Res. **8**(4), 419–430 (1982)

Marin, J., Biswas, A., Ofli, F., Hynes, N., Salvador, A., Aytar, Y., Weber, I., Torralba, A.: Recipe1m+: A dataset for learning cross-modal embeddings for cooking recipes and food images. IEEE Trans. Pattern Anal. Mach. Intell. **43**(1), 187–203 (2019)

Massimo, D., Elahi, M., Ge, M., Ricci, F.: Item contents good, user tags better: Empirical evaluation of a food recommender system. In: Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization. UMAP '17, pp. 373–374. Association for Computing Machinery, New York, NY, USA (2017). https://doi.org/10.1145/3079628.3079640

Mayfield, E., Black, A.W.: Should you fine-tune bert for automated essay scoring? In: Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 151–162 (2020)

Melville, P., Mooney, R.J., Nagarajan, R., et al.: Content-boosted collaborative filtering for improved recommendations. AAAI/IAAI **23**, 187–192 (2002)

Messina, P., Dominguez, V., Parra, D., Trattner, C., Soto, A.: Content-based artwork recommendation: integrating painting metadata with neural and manually-engineered visual features. User Model. User-Adapted Interact. **29**(2), 251–290 (2019). https://doi.org/10.1007/s11257-018-9206-9

Metzler, D., Bernstein, Y., Croft, W.B., Moffat, A., Zobel, J.: Similarity measures for tracking information flow. In: Proceedings of the 14th ACM International Conference on Information and Knowledge Management, pp. 517–524 (2005). https://doi.org/10.1145/1099554.1099695

Mosbach, M., Andriushchenko, M., Klakow, D.: On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines. arXiv preprint arXiv:2006.04884 (2020)

Mueller, J., Thyagarajan, A.: Siamese recurrent architectures for learning sentence similarity. In: Thirtieth AAAI Conference on Artificial Intelligence (2016). https://doi.org/10.1609/aaai.v30i1.10350

Musto, C., Trattner, C., Starke, A., Semeraro, G.: Towards a knowledge-aware food recommender system exploiting holistic user models. UMAP'20, pp. 333–337. Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3340631.3394880. https://doi.org/10.1145/3340631.3394880

Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: Bleu: A method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 311–318 (2002). https://doi.org/10.3115/1073083.1073135

Peng, S., Cui, H., Xie, N., Li, S., Zhang, J., Li, X.: Enhanced-rcnn: an efficient method for learning sentence similarity. In: Proceedings of The Web Conference 2020, pp. 2500–2506 (2020). https://doi.org/10.1145/3366423.3379998

Penha, G., Hauff, C.: What does bert know about books, movies and music? probing bert for conversational recommendation. In: Fourteenth ACM Conference on Recommender Systems, pp. 388–397 (2020)

Peska, L., Vojtas, P.: Off-Line vs. On-Line Evaluation of Recommender Systems in Small E-Commerce, pp. 291–300. Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3372923.3404781

Qin, Y., Wang, P., Li, C.: The world is binary: Contrastive learning for denoising next basket recommendation. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 859–868 (2021). https://doi.org/10.1145/3404835.3462836

Ren, P., Chen, Z., Li, J., Ren, Z., Ma, J., de Rijke, M.: Repeatnet: A repeat aware neural recommendation machine for session-based recommendation. Proceedings of the AAAI Conference on Artificial Intelligence **33**(01), 4806–4813 (2019). https://doi.org/10.1609/aaai.v33i01.33014806

Rendle, S., Freudenthaler, C., Gantner, Z., Schmidt-Thieme, L.: Bpr: Bayesian personalized ranking from implicit feedback. In: Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence, pp. 452–461 (2009)

Rendle, S., Freudenthaler, C., Schmidt-Thieme, L.: Factorizing personalized markov chains for next-basket recommendation. In: Proceedings of the 19th International Conference on World Wide Web, pp. 811–820 (2010). https://doi.org/10.1145/1772690.1772773

Ricci, F., Rokach, L., Shapira, B.: In: Ricci, F., Rokach, L., Shapira, B. (eds.) Recommender Systems: Introduction and Challenges, pp. 1–34. Springer, Boston, MA (2015). https://doi.org/10.1007/978-1-4899-7637-6_1

Rietzler, A., Stabinger, S., Opitz, P., Engl, S.: Adapt or get left behind: Domain adaptation through bert language model finetuning for aspect-target sentiment classification. In: Proceedings of the 12th Language Resources and Evaluation Conference, pp. 4933–4941 (2020)

Robertson, S., Walker, S., Jones, S., Hancock-Beaulieu, M.M., Gatford, M.: Okapi at trec-3. In: Overview of the Third Text REtrieval Conference (TREC-3), pp. 109–126. Gaithersburg, MD: NIST (1995). https://www.microsoft.com/en-us/research/publication/okapi-at-trec-3/

Rossetti, M., Stella, F., Zanker, M.: Contrasting offline and online results when evaluating recommendation algorithms. In: Proceedings of the 10th ACM Conference on Recommender Systems, pp. 31–34 (2016). https://doi.org/10.1145/2959100.2959176

Sánchez, P., Bellogín, A.: Attribute-based evaluation for recommender systems: Incorporating user and item attributes in evaluation metrics. In: Proceedings of the 13th ACM Conference on Recommender Systems, pp. 378–382 (2019). https://doi.org/10.1145/3298689.3347049

Sarwar, B., Karypis, G., Konstan, J., Riedl, J.: Item-based collaborative filtering recommendation algorithms. In: Proceedings of the 10th International Conference on World Wide Web, pp. 285–295 (2001). https://doi.org/10.1145/371920.372071

Schedl, M., Ferwerda, B.: Large-scale analysis of group-specific music genre taste from collaborative tags. In: 2017 IEEE International Symposium on Multimedia (ISM), pp. 479–482 (2017). IEEE

Sellam, T., Das, D., Parikh, A.: Bleurt: Learning robust metrics for text generation. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 7881–7892 (2020). https://doi.org/10.18653/v1/2020.acl-main.704

Shani, G., Gunawardana, A.: In: Ricci, F., Rokach, L., Shapira, B., Kantor, P.B. (eds.) Evaluating Recommendation Systems, pp. 257–297. Springer, Boston, MA (2011). https://doi.org/10.1007/978-0-387-85820-3_8

Shao, E., Guo, S., Pardos, Z.A.: Degree planning with plan-bert: Multi-semester recommendation using future courses of interest. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 14920–14929 (2021)

Sun, X., Meng, Y., Ao, X., Wu, F., Zhang, T., Li, J., Fan, C.: Sentence similarity based on contexts. Trans. Assoc. Comput. Linguist. **10**, 573–588 (2022). https://doi.org/10.1162/tacl_a_00477

Symeonidis, P., Janes, A., Chaltsev, D., Giuliani, P., Morandini, D., Unterhuber, A., Coba, L., Zanker, M.: Recommending the video to watch next: an offline and online evaluation at youtv.de. In: Fourteenth ACM Conference on Recommender Systems, pp. 299–308 (2020). https://doi.org/10.1145/3383313.3412257

Trattner, C., Elsweiler, D.: Investigating the healthiness of internet-sourced recipes: Implications for meal planning and recommender systems. In: Proceedings of the 26th International Conference on World Wide Web, pp. 489–498 (2017). https://doi.org/10.1145/3038912.3052573

Trattner, C., Jannach, D.: Learning to recommend similar items from human judgments. User Model. User-Adapted Interact. **30**(1), 1–49 (2020). https://doi.org/10.1007/s11257-019-09245-4

V Ganesan, A., Matero, M., Ravula, A.R., Vu, H., Schwartz, H.A.: Empirical evaluation of pre-trained transformers for human-level NLP: The role of sample size and dimensionality. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4515–4532. Association for Computational Linguistics, Online (2021). https://doi.org/10.18653/v1/2021.naacl-main.357. https://aclanthology.org/2021.naacl-main.357

Valcarce, D., Bellogín, A., Parapar, J., Castells, P.: On the robustness and discriminative power of information retrieval metrics for top-n recommendation. In: Proceedings of the 12th ACM Conference on Recommender Systems, pp. 260–268 (2018). https://doi.org/10.1145/3240323.3240347

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Proceedings of the 30th International Conference on Advances in Neural Information Processing Systems vol. 30, pp. 5998–6008 (2017)

Wan, M., Wang, D., Liu, J., Bennett, P., McAuley, J.: Representing and recommending shopping baskets with complementarity, compatibility and loyalty. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, pp. 1133–1142 (2018). https://doi.org/10.1145/3269206.3271786

Wang, P., Guo, J., Lan, Y., Xu, J., Wan, S., Cheng, X.: Learning hierarchical representation model for next basket recommendation. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR'15, pp. 403–412 (2015). https://doi.org/10.1145/2766462.2767694

Weber, I., Achananuparp, P.: Insights from machine-learned diet success prediction. Proceedings of Pacific Symposium on Biocomputing (PSB) **21**, 540–551 (2016). https://doi.org/10.1142/9789814749411_0049

Yao, Y., Harper, F.M.: Judging similarity: a user-centric study of related item recommendations. In: Proceedings of the 12th ACM Conference on Recommender Systems, pp. 288–296 (2018). https://doi.org/10.1145/3240323.3240351

Ying, H., Zhuang, F., Zhang, F., Liu, Y., Xu, G., Xie, X., Xiong, H., Wu, J.: Sequential recommender system based on hierarchical attention network. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence, pp. 3926–3932 (2018)

Yu, F., Liu, Q., Wu, S., Wang, L., Tan, T.: A dynamic recurrent model for next basket recommendation. In: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 729–732 (2016). https://doi.org/10.1145/2911451.2914683

Yu, L., Sun, L., Du, B., Liu, C., Xiong, H., Lv, W.: Predicting temporal sets with deep neural networks. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 1083–1091 (2020). https://doi.org/10.1145/3394486.3403152

Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with BERT. In: Proceedings of the 8th International Conference on Learning Representations (2020)

Zhang, T., Wu, F., Katiyar, A., Weinberger, K.Q., Artzi, Y.: Revisiting few-sample bert fine-tuning. arXiv preprint arXiv:2006.05987 (2020)

Zimdars, A., Chickering, D.M., Meek, C.: Using Temporal Data for Making Recommendations. In: Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence UAI'01, pp. 580–588 (2001)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Yue Liu received her Master's Degree from National University of Singapore in 2023. She is currently working at Singapore Management University. Her research interests include recommendation system and natural language processing.

Palakorn Achananuparp is a Senior Research Scientist at the School of Computing and Information Systems (SCIS), Singapore Management University. With a broad range of research interests, including Artificial Intelligence, Machine Learning, Natural Language Processing, Social Network Analysis, Computational Social Science, and Digital Health, he focuses on exploring the intersection of technology and society in various domains such as social media, politics, and public health, with the aim of leveraging computational methods to address complex societal challenges and improve well-being of individuals.

Dr. Ee-Peng Lim is the Lee Kong Chian Professor with the School of Computing and Information Systems at the Singapore Management University. He is also the Director of Living Analytics Research Centre in the School, a research centre focusing developing personalized and participatory analytics capabilities for smart city and smart nation relevant applications. Dr Lim received his PhD degree from University of Minnesota. His research expertise covers social media mining, social/urban data analytics, and information retrieval. He is the recipient of the Distinguished Contribution Award at the 2019 Pacific Asia Conference on Knowledge Discovery and Data Mining (PAKDD), and the Test of Time award at 2020 ACM Conference on Web Search and Data Mining (WSDM). He is currently a member of Singapore's Social Science Research Council.