



Exploring the Effect of Recipe Representation on Critique-Based Conversational Recommendation

Fakhri Abbas^(✉), Nadia Najjar, and David Wilson

University of North Carolina at Charlotte, Charlotte, NC 28223, USA
{fabbas1,nanajjar,davils}@uncc.edu

Abstract. Diet diversification can facilitate both positive health outcomes and greater enjoyment in food consumption. Case-based recommendation can play an important role to promote a more diverse diet by connecting users with more diverse options in meal planning. Our research investigates conversational CBR approaches to support greater diversity in recommendations for recipes and meal planning. This paper presents and evaluates a critique based conversational recommender approach to support diet diversification. The approach incorporates both (1) dynamic generation of diversity driven critique for conversational interaction, and (2) identification of key recipe features in contexts that promote greater diversity of results through dynamic critique. Our approach is evaluated using an initial offline simulation study, followed by a full online user study. Results show that meal diversity outcomes for users can be increased using dynamic critique, and that recipe representation plays an important role in the diversity of recommended recipes and has a direct impact on user choices and outcomes.

Keywords: Diversity · Recipe recommendation · Critique-based · Case-based reasoning

1 Introduction

Diet diversification has been linked to positive health outcomes such as reducing incidence of cancer or mortality [10]. However, due to a range of physiological, psychological, social and environmental factors, changing food-related behaviour such as adopting a diverse diet is a challenging task [6]. Hence, facilitating individual access to and exploration of diverse food choices is a step toward diet diversification. To help address this, our research investigates CBR recommender support for diversity during recipe exploration. Incorporating diversity in recipe recommenders provides a number of advantages. Diversity enables the user to explore alternative options that could be healthier and increase dietary diversity [10]. It also increases user awareness and knowledge of existing recipes by providing more recipes that could be explored from different cultures, cuisines, or communities [13]. But these advantages come with the challenge of balancing accuracy and diversity. The recommended diverse recipes must also meet user

requirements. In particular, finding the balance between diversity and accuracy is an open research challenge [13].

In previous work, we developed a CBR approach that enabled users compile diverse meal plans through the use of dynamic critique [1–3]. The research reported in this paper builds upon our previous work in critique generation, and expands upon it in two primary ways. First, we propose and evaluate a new model for critique generation to promote diversity. Second, we present an investigation of the impacts of different recipe case representations for critique generation and effectiveness. Our investigation addresses the following research questions (RQ) and hypotheses (H):

RQ 1: How does the proposed critique-based CBR approach impact the diversity of recommendations?

- **H 1.1** Critique-based recommendation results in recommending more diverse recipes compared to a non critique-based recommender.
- **H 1.2** Critique-based recommendation achieves higher diversity scores in fewer iterations compared to a non critique-based recommender.

RQ 2: In critique-based conversational recommendation, how does the underlying representation of recipe cases impact diversity in terms of user outcomes?

- **H 2.1** In diversity-focused critique, different recipe representations result in differences in the diversity of meal plans created by users.
- **H 2.2** In diversity-focused critique, recipe representations lead users to choose different types of critique features.
- **H 2.3** In diversity-focused critique, meal plan diversity is realized based on different features that are related to certain demographic characteristics.
- **H 2.4** In diversity-focused critique, recipe representation affects user perceptions of diversity.

The remainder of this paper is organized as follows. Section 2 presents an overview of related work. Section 3 presents our proposed approach for diversity-focused conversational recommendation. Section 4 presents our simulation study to address RQ 1, and the full user study evaluation to address RQ2. The paper concludes in Sect. 5 with discussion and future directions.

2 Background

This paper brings together three lines of related research: diversity in recommender systems, critiquing in conversational recommender systems, and the domain of recipe recommendation.

2.1 Diversity in Recommender Systems

The concept of diversity in recommender systems has been linked to the concept of similarity [25]. Smyth and McClave suggested measuring the diversity of recommended cases as the average pairwise distance [25]. Using pairwise distances

between cases to measure diversity has been widely adopted with variations in the distance metric (e.g., cosine metric, Jaccard similarity, etc.) [11, 19, 27, 29]. These differences in distance metrics depend on case representation. For example, when cases are represented by their content, the distance has been measured using the complement of Jaccard similarity [27], the complement of cosine similarity [11], or taxonomy-based metrics [29]. When cases are represented by rating, hamming distance [19], the complement of Pearson correlation [27], or the complement cosine similarity have been adopted as a distance measure. Our proposed approach adopts the average pairwise distance as a diversity measure.

2.2 Critique-Based Conversational Recommender Systems

Recommender systems are most often considered as a type of one shot interaction, in which the system recommends a set of items and the user navigates through that set to find an item of interest. Conversational CBR [4] and more generally conversational recommender systems (CRSs) take a different approach, providing a richer user interaction through iterative feedback and refinement of results. During successive iterations the system can elicit and refine the user's preference and context. This in turn has a positive impact on enabling users to better understand the search space, and reduce the effect of the cold start problem [18, 29]. For example, McGinty and Smyth [22] incorporated diversity in CRS while balancing the tradeoff between diversity and relevance. In each cycle of that approach the user selects a critique which is the basis for the next conversational step. The search is widened if the same critique is applied to the same case, and narrowed if a different critique is used on a different case. In [20], McCarthy et al. addressed diversity in critiquing, but the focus was on creating diversity in critiques rather than diversity in conversational outcomes.

Smyth and McGinty [26] compared different types of CRS. In particular, one promising approach employed a critiquing form of feedback. In critiquing feedback, the user provides a directional preference over a feature of the recommendation [9, 21]. For example, in a recipe recommender a user may ask for recipes with less meat and more protein. CRS in turn will adopt and recommend more vegetarian recipes with high protein. Here, meat and protein are the relevant recipe features, and less and more are the direction preferences. The feature(s) along with the direction(s) together comprise the critique.

Critiquing can be static or dynamic. Static critique is an approach in which there is a pre-designed set of potential critiques, which are fixed within the user interaction session. In contrast, the dynamic critique approach generates a unique set of potential critiques for each recommended item individually, based on a specified metric. An example dynamic critique approach was proposed by McCarthy et al. [21], in which the system combines features depending on the available items in the search space. Here, we investigate a novel dynamic critiquing approach to support diet diversification, which expands upon our initial work in case-based recommendation [1].

2.3 Diversity in Recipe Recommenders

Incorporating diversity into food recommender systems is a natural extension of Health-Aware recommenders. The importance of diversity in recipe recommendation has several advantages such as: providing meals with varied sources of nutrition for a balanced meal diet [12], increasing user awareness of existing recipes, and covering a wide variety of options that could reduce the cold start problem [5]. A number of studies have considered recommender support for making healthier meal choices. For example, Grace et al. [15] proposed a system (Q-Chef) that encourages dietary diversity by generating and recommending recipes based on models of surprise and novelty of the ingredients that appear in recipes. While Q-Chef focused on identifying new recipes that are surprising to the user and could result in diversifying their diet, the set of recommended recipes itself is not necessarily diverse. Similarly, Musto et al. [23] introduced a natural language justification approach to support food recommendation with the goal to promote healthy choices. This approach focused on transparent recommendation and not on diversity outcomes as such. Elswailer et al. [12] acknowledged the importance of diversity in meal plans as a way to provide healthy alternatives. They proposed a meal planner algorithm to recommend recipes, but acknowledged that diversity was not specifically engineered into the recommendations.

3 DiversityBite Framework: Recommend, Review, Revise

This research expands upon our *DiversityBite* CBR framework [1]. DiversityBite involves a three-stage *recommend-review-revise* cycle for CRS, as shown in Fig. 1. The *recommend stage*, consists of two main steps: recommending candidate cases and dynamically generating potential user critiques for each case. In the *review stage*, the user reviews the cases and can select a critique for one of them as the basis for a new set of recommended cases (or select a recipe as an outcome). Finally, in the *revise stage*, the selected critique serves as a constraint that is applied for the next recommend stage.

In practice, the cycle begins with a zooming phase [7] based on initial user context. The user context consists of both hard constraints (e.g., vegetarian) to filter out irrelevant recipes, and soft constraints (e.g., meal course) that provide relative weighting as part of retrieving baseline cases. The baseline recommendation step consists of applying a straightforward similarity metric to find recipes matching the user context. Potential critiques are then dynamically generated for each of the initial recommendations.

The dynamic critique generation step is the key part of the conversational process, and is the focus of our investigation. To support diversity in outcomes, a potential critique (e.g., more spicy) is presented as an option for the user only if it also promotes diversity in candidate recommendations. The critique is essentially “like this, but more diverse” across different recipe dimensions. Selecting diversity-positive critiques is accomplished by identifying a diversity goal—a representative subset of recipes that serves as a footprint for the space of

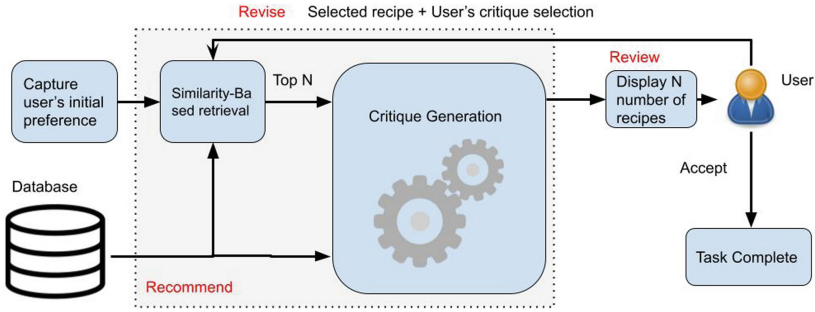


Fig. 1. Illustration of recommend-review-revise cycle found in DiversityBite along with the main components. The framework starts with user initial preference and ends with user acceptance. The shaded area represents retrieval and critique generation cycle.

available recipes. The baseline in [1] employed a stochastic process for generating the diversity goal, which we refer to as *Diversity Goal Footprint* (DGF).

3.1 Adaptive Diversity Goal Approach

In this research, we are proposing a new model for generating diversity goals, which we refer to as the *Adaptive Diversity Goal* (ADG) approach. Critique generation starts by identifying a diverse set of recipes within the domain of recipes that matches the user’s initial preference. This set represents a diversity goal, and it provides a basis for selecting critique features to move the recommendation toward a more diverse set of recipes. More specifically, the diversity goal is a set of recipes that serves as a reference point to select a critique and identify a direction (more/less) across the critique dimensions. The next step is to extract features related to critique from the diversity goal set, such as the average protein content in the set of recipes. The process then compares extracted diversity goal features with features found in a candidate recommended recipe. The comparison provides insight on which features should be forwarded as *critique features*, which, if selected, will increase diversity by moving the recommendation toward the diversity goal. The process concludes by identifying text for each critique feature along with the direction.

The proposed ADG approach establishes a diversity goal as a maximally diverse set of recipes from available recipes by applying a shortest path algorithm. This is a separate, contextually-dependent analysis in relation to each Top-N recipe case being suggested to the user. This addresses two limitations in the previous DGF approach. First, it considers the diversity analysis as an optimization problem that tries to find the optimal diverse set by utilizing a shortest path algorithm. Second, for each recommended recipe a new diversity goal is calculated. This process better aligns generated critique features to the proposed recipe cases.

Figure 2 shows a comparison between our previous DGF approach and the ADG approach proposed here. Figure 2 (left), compares between the new pro-

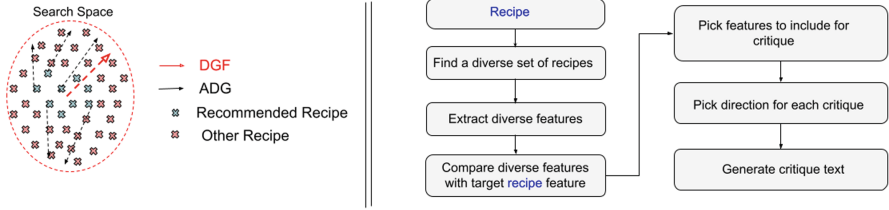


Fig. 2. A comparison between DGF vs. ADG (left). The general steps to generate critique proposed in [1] (right)

posed approach—Adaptive Diversity Goal (ADG) and the approach presented in [1]—Diversity Goal Footprint (DGF). It shows that each recipe will have a different diversity goal and this goal is guaranteed to be a maximally diverse set of recipes. Figure 2 (right) shows the common steps in generating critique proposed in this work. Section 4 will discuss the implementation details of ADG along with the algorithms used.

4 Evaluation

To evaluate our research questions, we conducted two primary evaluations. The first is an offline evaluation, a simulation study focused on addressing RQ 1. The second is an online evaluation, a full user study focused on addressing RQ 2. In the online evaluation, all user actions are logged for further analysis and exploration. This section first describes the case base used for experimentation. Next, it presents the implementation details for DGF, AGD, and diversity scoring, followed by the results of both the simulation study and the user study.

4.1 Case Base

Our experiments employ a recipe case library based on a dataset with strong potential for meal diversity—it contains a wide variety in terms of both recipes and cuisines. Sajadmanesh et al. [24] prepared a dataset with 120K recipes crawled from yummys.com, a personalized recipe recommender platform. The dataset consists of recipes from 204 countries. Each recipe has an average review rating, ingredients, preparation time, course type, nutritional values, and flavor features. The raw data contains 11,113 ingredients. The course type feature has values related to the recipe type such as afternoon tea, bread, breakfast, etc. The nutritional value features are: saturated fat, trans fat, fat, carbohydrate, sugar, calories, fiber, cholesterol, sodium, and protein of a recipe per serving. Recipes are characterized by six flavor features on a scale from 0 to 1: saltiness, sourness, sweetness, bitterness, spiciness, and savoriness. Preparatory analysis did not reveal substantive inconsistencies or errors in the feature data, so all features were included in the case base for evaluation. In addition, to reduce overall

ingredient sparsity we employed the FOODON [16] ontology to map each ingredient to a food concept. The mapping reduced the number of unique ingredients from 11,113 ingredients to 3,807 ingredients in case features.

4.2 Implementation: DGF, AGD, and Diversity Scoring

This section discusses comparative implementation detail for DGF and ADG in the evaluation studies. To create the diversity goal set using DGF [1], \mathbf{L} number of recipes were randomly selected from the search space and a diversity score would be calculated for that set. This process is repeated \mathbf{R} number of times. The list with the highest diversity score is then selected to represent the actual diversity goal. This approach follows Vargas et al. [28], who noted that maximum diversity can be approximated through random selection.

In ADG, the diversity goal is created using a greedy re-ranking algorithm, which employs Dijkstra’s shortest path analysis [8] to select the next recipe to be added to the list \mathbf{S} . In particular, given a current recipe \mathbf{r} then the next recipe \mathbf{n} should be the farthest (most dissimilar) recipe from \mathbf{r} within the search space. To help address nearest-neighbor search efficiency, a K-D tree algorithm [14] was employed to calculate the distances between recipes. In the ADG approach, each recipe in the recommended Top-N recipes serves as the start node of the shortest path algorithm to estimate the diversity goal.

The diversity score is calculated using the average pairwise distances between recipes following Smyth and McClave [25], as shown in Eq. 1.

$$Diversity(R) = \frac{\sum_{i \in R} \sum_{j \in R/\{i\}} dist(i, j)}{|R|(|R| - 1)} \quad (1)$$

where R , represents the recipe cases in the list.

4.3 Simulation Study: Incorporating Diversity in Critique

To address our first research question, a simulation study for critique generation was conducted. The purpose of the simulation study is (1) to understand the diversity scores of the recommended recipes over the course of the users’ interactions, and (2) the feasibility of the proposed algorithm by examining the change of diversity scores over the number of iterations. The following sections describe the experimental setup and evaluation results.

Experiment Setup. To evaluate the critique approach, three variations of DiversityBite were implemented, one without critique and two with critique. The first variation without critique (DiversityBite-) simulates a similarity-based recipe recommender. The second (with critique) variation implements the DGF (DiversityBite+DGF) approach, while the third (with critique) variation implements the ADG (DiversityBite+ADG) approach. In all variations the same DiversityBite *recommend-review-revise* cycle was employed; the only difference

is the variation in critique. The first iteration in the recommendation starts recommending the closest N recipes to the centroid of the user search space, where the centroid represents the average score for the ingredients vector. For each iteration, given a selected recipe from the previous iteration the algorithm selects the closest N recipes to the selected recipe. The closest N recipes were determined using the cosine similarity metric, where each recipe case is represented by a vector of 3,807 ingredients. After the selection of a recipe and a critique, the algorithm recommends the N closest recipe cases to the selected recipe with the critique applied.

The simulation consisted of building 100 user profiles. Each profile is evaluated by simulating 50 iterations of using DiversityBite-, DiversityBite+DGF, and DiversityBite+ADG. Since the yummly.com data does not provide user interaction with recipes, user profiles were created by randomly selecting a region. Then, a subset of recipes with an average rating of 4 or more (on a 1–5 scale) were randomly selected to build the user profile. The search space for each user is the rest of recipes found in the region but not in the user profile. To simulate user selection of recipes at each iteration, the closest recipe to the user profile centroid was chosen using cosine similarity. In the critique approach, a random critique was chosen from the critique list of the selected recipes.

For the implementation of the three variations the following settings were used: $N = 10$, $L = 100$. The total number of critique features is 16 (6 flavour + 10 nutrition). To ensure the reproducibility of the results, the user’s unique identifier was used as the random seed in the cases where randomness was used. Recipe cases were represented as a vector of ingredients with binary values where 1 means the ingredient is present in the recipe. Cosine similarity was used for recommendation while diversity calculation employed euclidean distance.

Analysis and Results. Results for RQ 1 addressed both diversity improvement (H 1.1) and number of iterations (H 1.2).

H1.1 - Diversity Improvement Analysis. The diversity score for the recommended set of recipes was measured at each iteration using the diversity Eq. 1. The left side of Fig. 3 shows the diversity score of the first 15 iterations for the same user in DiversityBite-, DiversityBite+DGF, and DiversityBite+ADG. In all iterations (except one) the diversity scores from DiversityBite+ADG were higher than the diversity score for DiversityBite- and DiversityBite+DGF. Lower scores are sometimes due to the simulator selecting a critique with a lower diversity score at one iteration compared to the rest. However, the overall score for other iterations show that DiversityBite+ADG is consistently higher. Figure 3 shows a comparison of diversity scores between the three recommender variations. The right side of Fig. 3 shows the overall distribution for each variation. To address the first hypothesis (H1.1), a one-way repeated measure ANOVA test shows there is a significant difference between the diversity scores in the recommended recipe for each iteration ($F(2,98) = 4.17$, $p < 0.05$). Tukey’s post hoc test shows that diversity in DiversityBite+DGF ($M = 2.27$, $SD = 0.37$), and DiversityBite+ADG ($M = 2.28$, $SD = 0.37$) is significantly higher than the diversity

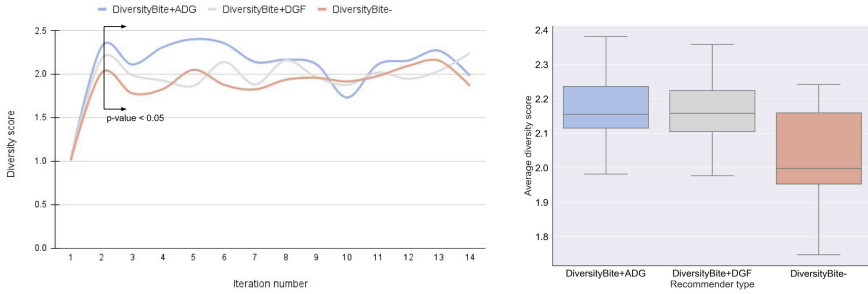


Fig. 3. Comparison between DiversityBite-, DiversityBite+DGF, and DiversityBite+ADG diversity scores

scores in DiversityBite- ($M = 2.08$, $SD = 0.39$). This analysis indicates that using critique has the potential to enable recommendation of more diverse recipes. While in simulation we did not find a statistically significant difference between DiversityBite+DGF, and DiversityBite+ADG, results indicate that both diversity critique methods can be used to increase diversity compared to the baseline DiversityBite-.

H 1.2 - Diversity Improvement and Number of Iterations. To address the second hypothesis H1.2, we investigated the relation between the number of iterations and the diversity scores. Our analysis shows that from the second iteration there is a statistically significant difference ($p < 0.05$) between the critique-based recommenders (DiversityBite+DGF, DiversityBite+ADG) and the non-critique recommender (DiversityBite-), as shown in Fig. 3. The results at each iteration support H1.2, with higher diversity in the critique-based approach. This indicates that the critique-based approach can be applied in real scenarios even with comparatively few critique interactions.

4.4 User Study: Comparing Different Recipe Representations

To address the second research question, a full version of DiversityBite was developed and deployed to conduct a user study, in which users were asked to prepare a weekly meal plan by interacting with the system to explore recipes.

Experiment Setup. A web-based recommender application of DiversityBite was developed for users to interact with. Figure 4 shows a screenshot of the web application used in the study. The user study analyzed performance of the approach and compared the impact of different types of underlying recipe representation for dynamic critique. For this study, the Adaptive Diversity Goal approach was employed. Approach parameters were set for Top-N recommendation $N = 10$ (# recommended recipe cases), and ADG $S = 10$ (cardinality for diversity goal set). The parameters were selected based on pilot testing to ensure reasonable computation time during user interaction with the website.

The screenshot displays a recipe interface with the following components:

- Recipe Card 1:**
 - Name: Hearty Vegetarian Chili with Butternut Squash
 - Ingredients: olive oil, onions, chile pepper, carrots, celery, jalapeno chilies, garlic, chili powder, paprika, ground cumin, dried oregano, salt, pepper, tomatillos, vegetable broth, worcestershire sauce, butternut squash, yukon gold potatoes, white beans, pinto beans, frozen corn, fresh cilantro, lime juice, aged cheddar cheese, sour cream, radishes, lime wedges, tortilla chips.
 - Preparation time: 95.0 minutes
- Recipe Card 2:**
 - Name: Zucchini and Corn Chili
 - Ingredients: poblano chilies, olive oil, frozen zucchini, onions, garlic, chili powder, cumin, low sodium vegetable stock, diced tomato nectar, cilantro, lime.
 - Preparation time: 40.0 minutes
- Recipe Card 3:**
 - Name: Vegetarian Sweet Potato & Chili
 - Ingredients: olive oil, sweet potatoes, onion, pepper, garlic cloves, tomato paste, chile p, cumin, kosher salt, freshly ground black pepper, tomatoes, black beans, honey, orange, tortilla chips, avocado, goat cheese, green onions, jalapeno chilies, radishes.
 - Preparation time: 70.0 minutes
- Nutrition features:**
 - CALORIES: 250.0
 - SODIUM: 720.0mg
 - Fat: 5.0g
 - Protein: 12.0g
 - Carbs: 43g
 - 23% DV
 - 14%
- Flavor features:**
 - Spiciness: High (red bar)
 - Sourness: Medium (orange bar)
 - Saltiness: Low (blue bar)
 - Sweetness: Very Low (green bar)
 - Bitterness: Very Low (yellow bar)
 - Meatness: Low (dark blue bar)
- Critique options:**
 - Less Spicy (red arrow down)
 - Less Sour (red arrow down)
 - Less Salty (red arrow down)
 - More Carbohydrate (green arrow up)
 - More Sugar (green arrow up)
 - More Calories (green arrow up)

Fig. 4. A screenshot shows the interface of the web application. Participants can see details about the recipe including ingredients, flavor, and nutritional info. The explore more link displays the critique features the user can select to load more recipes.

Four variations of DiversityBite were implemented with four different representations: ingredient (*Ingr-DiversityBite*), flavor (*F-DiversityBite*), nutrition (*N-DiversityBite*), and flavor & nutrition (*FN-DiversityBite*) features. The baseline was taken as *Ingr-DiversityBite*, since it has been used in the simulation study while the other three variations are the treatment. In all variations, the user explores more recipes by using critique. Flavor and nutritional features were used as a critique for each variation regardless of the representation.

Participants. One-hundred participants were recruited from students, staff and faculty at a U.S. public university. Total time spent by each participant was on average 30 min. Participants included 67 females and 33 males, with the majority of participants in the age range of 18 to 24 years old. The majority of participants had at least a bachelor's degree. All participants reported using online resources to look for new recipes or to refresh their memory regarding a recipe they know. Additionally, all participants indicated that they frequently look for new recipes. Participants reported the most frequently used online resources as: Google search, YouTube videos, and social networks. The main criteria reported as considerations in seeking recipes were: recipe ingredients, preparation time, and balanced dish. This suggests that participants had a good exposure to online resources when looking for recipes. Among the chosen cuisines, Italian, American, Mexican, and Indian were the most frequently chosen cuisines, while the least chosen cuisines were Ethiopian, Swedish, and Ukrainian. This would seem to align with the demographic distribution of the area participants were recruited from (a U.S. urban public university). Since the task was to develop a weekly meal plan, the most frequently chosen meal courses were: main dish, Lunch, and Breakfast/Brunch, while the least frequently chosen ones are beverages such as

tea, and cocktail. On average participants spent around 5 min using each DiversityBite variation, and viewed on average 7 different recipe lists in each variation; so, they were presented with at least 70 different recipes in each variation.

Diversity in Meal Plan vs Diversity in Recommended Recipes. In order to analyze diversity in meal plans, five different dimensions of diversity in a meal plan were considered. The feature variations on case representation considered are: Ingredients (only), Flavor (only), Nutrition (only), Nutrition & Flavor (combined), and finally Ingredients & Flavor & Nutrition (combined).

Intuitively, the diversity of the meal plan depends on the diversity of the recommended recipes as participants created their meal plan from recommended recipes. Results of Pearson correlation analysis showed a significant direct relation between both of them, regardless of the diversity type or the case representation type. In terms of meal plan diversity, participants were able to create a more diverse meal plan in each variation. Table 1 summarizes the average diversity meal plan for each variation along with diversity definition. The Table shows the results in groups depending on the diversity definition, rows with the highest diversity scores highlighted. For example, using ingredient representation for recipes *Ingr-DiversityBite* participants created meal plans with highest diversity in ingredients. Similarly, participants created meal plan with high diversity in nutrition using *Ingr-DiversityBite* variation. The table also shows that, using *N-DiversityBite* variation participants created diverse meal plan in terms of flavor, and in the case of all features combined. Finally, using *F-DiversityBite* the meal plan created were diverse in terms of both nutrition and flavor.

Results show that different representation choices yield differences in meal plan diversity. For the Diversity-Ingr variation, a one-way repeated measure ANOVA test shows a significant difference between the diversity scores in the meal plan for each variation ($F(3,297) = 6.82, p < 0.05$). Tukey's post hoc test shows that diversity score in *Ingr-DiversityBite*, is significantly higher than the diversity scores in *F-DiversityBite*. For Diversity-N, a one-way repeated measure ANOVA test shows a significant difference between the diversity scores in the meal plan for each variation ($F(3,297) = 3.62, p < 0.05$). Tukey's post hoc test shows that diversity score in *Ingr-DiversityBite*, is significantly higher than the diversity scores in *F-DiversityBite*. For Diversity-NF, a one-way repeated measure ANOVA test shows a significant difference between the diversity scores in the meal plan for each variation ($F(3,297) = 4.36, p < 0.05$). Tukey's post hoc test shows that diversity score in *F-DiversityBite*, is significantly higher than the diversity scores in *FN-DiversityBite*. Finally, for all features combined, a one-way repeated measure ANOVA test shows there's a significant difference between the diversity scores in the meal plan for each variation ($F(3,297) = 5.64, p < 0.05$). Tukey's post hoc test shows that diversity score in *N-DiversityBite*, is significantly higher than the diversity scores in *Ingr-DiversityBite*. Despite the similarity and the high correlation between flavor, nutrition, and ingredient, overall results show that diversity in meal plan differ depending on the representation which supports hypothesis H2.1.

Table 1. Average meal plan diversity for each variation along with the definition of diversity

Diversity Type	Variation	Average Diversity	Diversity Type	Variation	Average Diversity
Ingredients	<i>Ingr-DiversityBite</i>	0.0951	Nutrition	<i>Ingr-DiversityBite</i>	0.3942
Ingredients	<i>F-DiversityBite</i>	0.0819	Nutrition	<i>F-DiversityBite</i>	0.3817
Ingredients	<i>N-DiversityBite</i>	0.0915	Nutrition	<i>N-DiversityBite</i>	0.3917
Ingredients	<i>FN-DiversityBite</i>	0.0930	Nutrition	<i>FN-DiversityBite</i>	0.3872
Flavor	<i>Ingr-DiversityBite</i>	0.3937	Nutrition+Flavor	<i>Ingr-DiversityBite</i>	0.3759
Flavor	<i>F-DiversityBite</i>	0.3955	Nutrition+Flavor	<i>F-DiversityBite</i>	0.3795
Flavor	<i>N-DiversityBite</i>	0.3957	Nutrition+Flavor	<i>N-DiversityBite</i>	0.3724
Flavor	<i>FN-DiversityBite</i>	0.3944	Nutrition+Flavor	<i>FN-DiversityBite</i>	0.3666
All features	<i>Ingr-DiversityBite</i>	0.1248	All features	<i>F-DiversityBite</i>	0.1344
All features	<i>N-DiversityBite</i>	0.1348	All features	<i>FN-DiversityBite</i>	0.1341

User Behaviour in Critique Selection. To address hypothesis H2.2, we studied participants’ behavior in the selection of different types of critique. Critiques can be viewed as two main types: Nutrition and Flavor. Nutrition critiques include: Protein, Calories, Carbohydrate, Sugar, Fiber, and Fat. Flavor critiques include: Bitter, Sour, Salty, Meaty, Spicy, and Sweet. Figure 5, shows that participants preferred to explore using flavor critique over nutrition critique in all variations except in flavor representation (*F-DiversityBite*). This suggests that *F-DiversityBite* was able to provide participants recipes with flavors that matched their preferences, and therefore participants chose nutrition critique to explore more. Result differences were statistically significant ($p < 0.05$) across all variations, with the exception of the flavor variation. These results support hypothesis H2.2 in which representation types can lead users to prefer one type of critique over the other. The *F-DiversityBite* variation recommended recipes that matched participants’ flavor interest but not nutrition. Therefore, participants chose to explore recipes using the nutrition critique feature. These results are also corroborated through participant responses in the reflection survey. The survey asked “What influenced your main decision when you explored more recipes?” and participants chose flavor as the main reason that influenced their decision to explore.

Demographic Differences in Diversity and Critique Selection. To address hypothesis H2.3, participant data was grouped and analyzed based on demographic categories collected in the survey. We asked participants for their age, gender, and education level. The results show no differences between the different groups in terms of their age and education when it comes to critique selection. However, the analysis showed differences among participants when grouped based on gender. In terms of meal plan diversity, male partici-

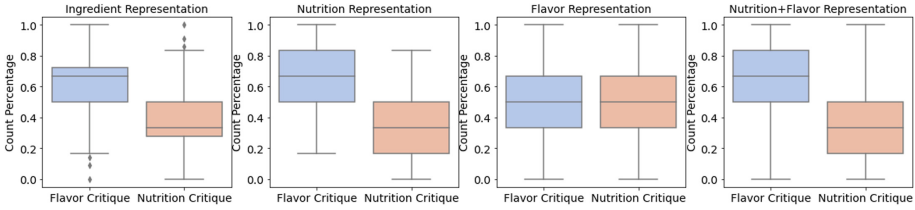


Fig. 5. A comparison between participants selection on flavor and nutrition critique on different variations

pants created more diverse meal plans using the nutrition representation, while female participants created more diverse meal plans using the flavor representation. One-way ANOVA with repeated measures shows a statistical significant in flavor and nutrition diversity scores (NF-Diversity) in meal plans for females ($F(3,198) = 6.25, p < 0.05$). Tukey’s post hoc test shows that diversity scores for the *F-DiversityBite* variation are significantly higher than the diversity scores in *FN-DiversityBite*. For males, a one-way ANOVA with repeated measures shows a statistically significant difference in nutrition diversity scores (N-Diversity) in meal plans ($F(3,96) = 3.70, p < 0.05$). Tukey’s post hoc test shows that diversity scores for *N-DiversityBite* variation are significantly higher in terms of diversity scores than *F-DiversityBite*. This suggests that male participants were more interested in diversifying the meal plan in terms of nutrition, while female participants were interested in diversifying meal plan in terms of flavor. To further explore this observation, critique selections for males and females were analyzed. The analysis shows that females chose to explore more recipes using flavor critiques compared to nutrition critiques. On the other hand, male participants used flavor and nutrition critiques across two representations equally. This finding supports hypothesis H2.3 in which demographic differences show different behavior in terms of meal plan diversity and critique behavior selection.

User Perceptions. To capture participant perceptions of diversity, participants were asked questions that address the variety of the recommended recipes and the created meal plan. According to [17], there are two types of diversity *categorical diversity* and *item-to-item diversity*. In the recipe domain, categorical diversity refers to diversity in terms of cuisine while item-to-item diversity refers to differences between the recipes such as ingredient, nutrition, and flavor. In the three questions, the term variety and different were used instead of diversity to avoid priming users specifically about diversity. Q1 (“I have seen recipes of different variety”), and Q2 (“I was able to create a meal plan from different variety of recipes”) address categorical diversity. Q3 (“Recipes in my meal plan were similar to each other”) addresses item-to-item diversity. The average rating for the questions Q1, Q2, and Q3 were 3.1, 3.2, and 3.7, respectively. In all variations none of the questions showed statistically significant differences, which aligns with previous findings [17]. Therefore, even though introducing diversity in rec-

ommendations while exploring shows a positive influence on diversity outcomes, it may not make a noticeable difference to participants' perceptions.

5 Conclusion

This paper presented a new approach and evaluation for generating dynamic critique to increase diversity in conversational case-based meal selection. Results of our initial simulation study confirmed that diversity can be increased using the proposed critique generation technique. Results of our user study showed that recipe representation has an effect on the diversity of a meal plan created by participants. In addition, different recipe case representations can lead users to select different types of critique. These two findings were also extended to show behaviours between different demographic groups. While this provides positive support for user outcomes, interestingly user perception of diversity was not significantly different between different recipe representations.

References

1. Abbas, F., Najjar, N., Wilson, D.: The bites eclectic: critique-based conversational recommendation for diversity-focused meal planning. In: Sánchez-Ruiz, A.A., Floyd, M.W. (eds.) ICCBR 2021. LNCS (LNAI), vol. 12877, pp. 1–16. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-86957-1_1
2. Abbas, F., Najjar, N., Wilson, D.: Increasing diversity through dynamic critique in conversational recipe recommendations. In: Proceedings of the 13th International Workshop on Multimedia for Cooking and Eating Activities (2021)
3. Abbas, F.G., Najjar, N., Wilson, D.: Critique generation to increase diversity in conversational recipe recommender system. In: The International FLAIRS Conference Proceedings (2021)
4. Aha, D.W., McSherry, D., Yang, Q.: Advances in conversational case-based reasoning. *Knowl. Eng. Rev.* **20**(3), 247–254 (2005)
5. Anderson, C.: A survey of food recommenders. arXiv preprint [arXiv:1809.02862](https://arxiv.org/abs/1809.02862)
6. Braude, H.D.: *Intuition in Medicine: A Philosophical Defense of Clinical Reasoning*. University of Chicago Press (2012)
7. Burke, R.D., et al.: Knowledge-based navigation of complex information spaces. In: Proceedings of the National Conference on Artificial Intelligence (1996)
8. Chen, J.C.: Dijkstra's shortest path algorithm. *J. Formalized Math.* **15**, 237–247 (2003)
9. Chen, L., Pu, P.: Critiquing-based recommenders: survey and emerging trends. *User Model. User-Adap. Interact.* **22**, 125–150 (2012)
10. Drescher, L.S., et al.: A new index to measure healthy food diversity better reflects a healthy diet than traditional measures. *J. Nutr.* **137**, 647–651 (2007)
11. Ekstrand, M.D., et al.: User perception of differences in recommender algorithms. In: Proceedings of the 8th ACM Conference on Recommender Systems (2014)
12. Elswiler, D., et al.: Bringing the “healthy” into food recommenders. In: DMRS (2015)
13. Freyne, J., Berkovsky, S.: Intelligent food planning: personalized recipe recommendation. In: Proceedings of the 15th International Conference on Intelligent User Interfaces (2010)

14. Goldberger, J., et al.: Neighbourhood components analysis. In: *Advances in Neural Information Processing Systems* (2004)
15. Grace, K., Maher, M.L., Wilson, D., Najjar, N.: Personalised specific curiosity for computational design systems. In: Gero, J.S. (ed.) *Design Computing and Cognition '16*, pp. 593–610. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-44989-0_32
16. Griffiths, E.J., et al.: FoodON: a global farm-to-fork food ontology. In: *ICBO/BioCreative* (2016)
17. Hu, R., Pu, P.: Helping users perceive recommendation diversity. In: *DiveRS@ RecSys* (2011)
18. Jannach, D., Manzoor, A., Cai, W., Chen, L.: A survey on conversational recommender systems. arXiv preprint [arXiv:2004.00646](https://arxiv.org/abs/2004.00646) (2020)
19. Kelly, J.P., Bridge, D.: Enhancing the diversity of conversational collaborative recommendations: a comparison. *Artif. Intell. Rev.* **25**, 79–95 (2006)
20. McCarthy, K., Reilly, J., Smyth, B., McGinty, L.: Generating diverse compound critiques. *Artif. Intell. Rev.* **24**, 339–357 (2005)
21. McCarthy, K., et al.: An analysis of critique diversity in case-based recommendation. In: *FLAIRS Conference* (2005)
22. McGinty, L., Smyth, B.: On the role of diversity in conversational recommender systems. In: Ashley, K.D., Bridge, D.G. (eds.) *ICCBR 2003*. LNCS (LNAI), vol. 2689, pp. 276–290. Springer, Heidelberg (2003). https://doi.org/10.1007/3-540-45006-8_23
23. Musto, C., et al.: Exploring the effects of natural language justifications in food recommender systems. In: *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization* (2021)
24. Sajadmanesh, S., et al.: Kissing cuisines: exploring worldwide culinary habits on the web. In: *Proceedings of the 26th International Conference on WWW Companion* (2017)
25. Smyth, B., McClave, P.: Similarity vs. diversity. In: Aha, D.W., Watson, I. (eds.) *ICCBR 2001*. LNCS (LNAI), vol. 2080, pp. 347–361. Springer, Heidelberg (2001). https://doi.org/10.1007/3-540-44593-5_25
26. Smyth, B., McGinty, L.: An analysis of feedback strategies in conversational recommenders. In: *The Fourteenth Irish Artificial Intelligence and Cognitive Science Conference (AICS 2003)* (2003)
27. Vargas, S., Castells, P.: Rank and relevance in novelty and diversity metrics for recommender systems. In: *Proceedings of the 5th ACM Conference on Recommender Systems* (2011)
28. Vargas, S., et al.: Coverage, redundancy and size-awareness in genre diversity for recommender systems. In: *Proceedings of the 8th ACM Conference on Recommender Systems* (2014)
29. Ziegler, C.N., et al.: Improving recommendation lists through topic diversification. In: *Proceedings of the 14th International Conference on World Wide Web* (2005)