IMPROVING DIVERSITY IN CONVERSATIONAL RECIPE RECOMMENDER THROUGH DYNAMIC CRITIQUING

by

Fakhri Abbas

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Computing and Information Systems

Charlotte

2022

Approved by:

____

Dr. David Wilson

____

Dr. Nadia Najjar

____

Dr. Mary Lou Maher

____

Dr. Kazjon Grace

____

Dr. Jing Yang

ABSTRACT

FAKHRI ABBAS. Improving Diversity in Conversational Recipe Recommender through Dynamic Critiquing. (Under the direction of DR. DAVID WILSON and DR. NADIA NAJJAR)

Diet diversification has been shown both to improve nutritional health outcomes and to promote greater enjoyment in food consumption. Conversational Recommender Systems (CRS) have a rich history in direct recommendation of recipes and meal planning, as well as conversational exploration of the possibilities for new food items. However, limited attention has been given to incorporating diversity outcomes as a primary factor in conversational critique for exploration. Critiquing as a method of feedback in recommendation has proven effective for conversational interactions, and diversifying recommended items during the exploration can help users broaden their food options, which critiquing alone may not achieve. All of these aspects together are important elements for recommender applications in the food domain.

This dissertation explores incorporating diversity in a critique-based conversational recommender system to support diet diversification. Recommender systems are known to support the task of exploitation while diversity supports the task of exploration. The research in this dissertation employs a conversational recommender approach to help maintain this balance — enabling exploration through critiquing and exploitation by selecting the closest matching recommendations to the user profile. To enable this balance this dissertation introduces an interactive critique-based conversational recipe recommender approach called *DiversityBite*, a novel way of dynamically generating critique during recipe recommendation.

This dissertation presents three studies — one simulation study and two user studies — to show the potential of using dynamic critique in increasing diversity. These studies investigate how the proposed *DiversityBite* approach can improve diversity in recipe recommendation. The contributions of this dissertation are: (i) Development and evaluation of a novel approach of dynamic diversity-focused critique for conversational recommender system, (ii) Applying dynamic diversity-focused critique in recipes domain to support diet diversification while exploring, and (iii) Identification of recipe features that are helpful in finding diverse recipes using dynamic critique. The results show that diversity can be increased using conversational recommender using dynamic-focused critique.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude and thanks to my advisors Dr. David Wilson, and Dr. Nadia Najjar. Their dedication and interest had been solely and mainly responsible for completing this dissertation. Thanks for your guidance and for allowing me to grow as a researcher.

Moreover, I would like to thank all the respected members in my dissertation committee: Dr. Mary Lou Maher, and Dr. Kazjon Grace for their valuable feedback that have enhanced the quality of my dissertation. Thanks for your insightful and instructive guidance that made this research so much better. A special thank you to Dr. Jing Yang for taking on the role of the Faculty Graduate Representative on my dissertation defense committee.

My greatest appreciation and thanks go to my parents, Ghassan and Sanaa Abbas, for their endless support. Thanks to my siblings Hanadi, Mai, and Mohammed for their encouragement.

Finally, I would like to extend my thanks to my friends, and lab mates, who have provided friendship and support through this journey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

Diet diversification has been linked to positive health outcomes such as reducing the incidence of cancer or mortality [4]. Moreover, by promoting a variety of healthful food choices, diet diversification can make food consumption more enjoyable [5]. In addition, nutritionists believe that variety in food is the key to an optimal diet [6]. However, due to a range of physiological, psychological, social, and environmental factors changing food-related behavior such as adopting a diverse diet is a challenging task [6]. For example, nutritional knowledge has been considered necessary but not sufficient for changing food behavior [7], and the accessibility for relevant information related to diet diversification is limited, too abstract, or contradicting [8]. Hence, providing individuals access to diverse food choices to explore is a step toward diet diversification.

In terms of making food choices, increasing numbers of online sites have emerged over the last decade that enable users to upload and share recipes (e.g., food.com, allrecipes.com). With more users sharing more recipes from different backgrounds, geographic locations, and cultures, there is a vast availability in terms of numbers of recipes, and diversity of cuisines. This has resulted in an overwhelming number of recipes for individuals to select from, which can make the search for a suitable recipe time-consuming. Therefore, there may be many suitable recipes that the user is not able to see or connect with. To address these issues, several recommender system approaches have been proposed for finding relevant recipes in a reasonable amount of time. For example, Freyne et. al. [9] created a recipe recommender that utilizes user ratings to find suitable recipes by looking at recipe ingredients. Despite the variety of available recipe choices, recommender systems have been criticized for providing a very myopic view of the world [10]. In particular, recommender systems helped users to find some suitable recipes in a reasonable amount of time, but that in turn affected diet diversification by providing only a limited number of overall food choices [9].

1.1 Motivation

Incorporating diversity in recipe recommenders provides several advantages. Diversity enables the user to explore alternative options that could be healthier, which can increase dietary diversity for individuals [4]. It also increases user awareness and knowledge of potential recipes by helping users to explore more recipes from different cultures, cuisines, or communities [9]. The problem of incorporating diversity mechanisms in recipe recommendation to generate diverse results that also meet user requirements remains an important open research challenge [9].

Primary aspects of the challenge stem from the differences and similarities between recipes (e.g., cuisine, ingredients, nutrition, meal-type, preparation), as well as users' perceptions of those differences and similarities. For example, a person may like fried chicken but not grilled chicken, or may prefer chicken for dinner but not for lunch. Due to this complexity in food selection, recommender systems should enable the user to shape the direction of the recommendations. This, in turn, helps to reduce the effect of contextual factors that are hard to capture such as time, cultural background, food knowledge, and the current user's needs.

Recommender systems typically monitor user interactions over time and tailor recommendations to user interests. Although this general approach is common and can be useful, there are several potential limitations [11]. First, the recommender system may not be able to estimate a user's preference, especially in early iterations i.e cold start problem. Second, recommendations may be highly context-dependent, such as in the domain of recipe recommendation. Finally, the user's knowledge about the domain can be limited and preferences may evolve over time [12]. Although monitoring users' interactions over time might be suitable to find recipes over an extended period of time, it depends on developing extensive user profiles, which is not possible in many cases. In contrast, this dissertation focuses on a conversational recommender approach, in which the user can provide iterative feedback on recommendations. The iterative feedback helps to reduce the effect of contextual factors that are hard to capture such as time, cultural background, food knowledge, and current user's needs.

In conversational-based recommenders, the type of feedback is also as important as the ability to provide feedback [1]. Conversational recommenders enables the user to guide the refinement of subsequent results to be explored. Smyth and McGinty [13] noted four main forms of conversational feedback. This research adopts their comparison scheme and employs critiquing as the primary form of user feedback. In critiquing, the user selects one or more features to refine the results for the next iterations. The selection of a critique enables the user to either narrow down or widen the range of possible results. Key research questions arise in generating potential critique interactions, such as selecting appropriate enabling features and critique options within the overall critique space.

1.2 Research Focus

This dissertation explores incorporating diversity in a critique-based conversational recommender system to support diet diversification. Recommender systems have been widely known to support the task of exploitation by tailoring recommendations to the user's needs. On the other hand, diversity has been used to increase exploration [14]. Combining the exploitation through the use of the baseline recommender approach and exploration through the use of diversity preserves a balance between the two sides [15]. Using a conversational recommender, this dissertation maintains this balance by enabling the exploration through critiquing and maintains the exploitation by selecting the closest recommendations to the user profile.

This dissertation introduces an interactive critique-based conversational recipe recommender approach called DiversityBite, a novel way of dynamically generating critiques, in which the generated critique leads the user to a more diverse set of recipe recommendations - and outcomes. In essence, this can be thought of as a "like this, but more diverse" approach. This interaction aims to promote diet diversification by keeping the balance of exploration and exploitation. In contrast to previous recipe recommender approaches, DiversityBite enables users to lead the interaction through the use of critique feedback. The critiques are generated dynamically to allow users to explore diverse recipes within the overall space of available recipes. To address the problem of incorporating diversity in recipe recommendation through dynamic critiquing, this dissertation addresses the following research questions:

- **RQ 1**: How do critique-based approaches impact the diversity of recommendations?
  - **H 1.1** Critique-based recommenders result in recommending more diverse recipes compared to non critique-based recommenders.
  - **H 1.2** Critique-based recommenders achieve higher diversity scores in less number of iterations compared to non critique-based recommenders.
- **RQ 2**: In critique-based conversational recommendation, how does diversity-focused critique impact diversity in terms of user outcomes?
  - **H 2.1** Critique-based recommenders result in finding more diverse recipes compared to a non critique-based recommenders.
  - **H 2.2** Users compile a more diverse meal plan in a critique-based recommender compared to non critique-based recommender.
  - **H 2.3** Users perceive the modeled diversity of the recommended recipes.
  - **H 2.4** In critique-based recommenders users prefer to explore using specific critique groups related to recipe features.

- **RQ 3**: In critique-based conversational recommendation, how does the underlying representation of recipes impact diversity in terms of user outcomes?
  - **H 3.1** In diversity-focused critique, different recipe representation results in differences in the diversity of meal plans created by users.
  - **H 3.2** In diversity-focused critique, recipe representation leads users to choose different types of critique features.
  - **H 3.3** In diversity-focused critique, the diversity of the meal plan is realized based on different features that are related to certain demographic characteristics.
  - **H 3.4** In diversity-focused critique, recipe representation impacts the user's perception of diversity.

1.3 Methods and Evaluation

This dissertation adopts a mixed-method approach to answer the previously stated research questions. This approach consists of three main activities. First, reviewing the literature to develop the research framework and highlight the importance of using critique, along with its relation to enabling exploration through user interaction. Second, designing and developing a critique-based CRS in the recipe domain to increase diversity in recommended recipes. Third, conducting a series of experiments to validate the hypotheses related to the research questions. This dissertation presents two types of experiments, offline experiments, and online experiments. Offline experiments are conducted to establish foundational results for a more comprehensive study. Online experiments are full user studies, where participants interact with a fully developed system.

1.4 Contributions

The contributions of this dissertation are as follows:

- Development and evaluation of a novel approach of dynamic diversity-focused critique for conversational recommender systems.

- Applying dynamic diversity-focused critique in recipes domain to support diet diversification while exploring.

- Identification of recipe features that help find diverse recipes using dynamic critique.

First, a novel approach of dynamic critique is proposed. Rather than presenting all critiques for each item, specialized critiques are generated dynamically. The dynamic critique approach is diversity-driven, in which the selected critiques were selected to increase the overall diversity of recommended items. This can be thought of as a "like this, but more diverse". Unlike other critique-based CRS that focus on increased similarity of recipes to fine-tune results, this dissertation employs critique to guide users into more exploration. This dissertation introduces two computational methods to select a subset of critiques for conversational feedback on recommended items.

The second contribution is related to the application domain. This work uses recipes as an application domain to apply the concept of dynamic critique in CRS. Increasing diversity in recipe recommenders is an ongoing research field that has proven to benefit users in the short and long run. Despite the benefits of using CRS in the recipes domain such as understanding the relationship between recipes, there has been comparatively little investigation of CRS as a way to help users navigate through the search space of recipes.

The last contribution is related to the domain of recipes. This dissertation tries to understand the relation between recipes' features and user exploration. Recipes can be represented using several different features such as ingredients, user rating, or flavor. Different underlying representations can lead to different results, which can affect the diversity of recommended recipes and therefore the exploration experience.

In this work, the effect of recipe representation on diversity is studied in the CRS.

1.5 Thesis Organization

This dissertation is organized as follows: Chapter 1 presents the purpose of the study, the research questions to be investigated, and the significance of the study. Chapter 2 reviews related work including conversational recommender systems, diversity in recommender systems, and food computing. Then, Chapter 3 presents the proposed overall CRS approach along with the models to generate dynamic critique. Chapter 4 presents three studies that are conducted to evaluate the proposed model and answer the research questions. Finally, Chapter 5 contains the discussion, limitations in the current work, and future directions for research.

CHAPTER 2: Overview of Related Research

This work brings together three lines of relevant research: conversational recommender systems, recipe recommendation, and diversity in recommender systems.

2.1 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. Typically, the system monitors users' interactions over time and tailors the recommendations to the user's interest. Although this approach is helpful and very common in many domains, it faces several limitations [11]. First, in some scenarios user preference cannot be reliably estimated. This can be the case in products that involve the customer in taking time before deciding in purchase (high involvement product) such as mobiles or laptops products, where there are no past observations. Second, the recommendation is highly context-dependent and it might be challenging to infer the user's needs and interests automatically. Finally, in some cases, an assumption was made that users know their preferences when they interact with the recommender system. However, this may not be the case. For example, users might construct their preferences during the decision process [12], and in other scenarios, users may learn about the domain and available options only during interaction with the recommender [16].

A conversational recommender system (CRS) can help to address some of these problems. CRS takes a different approach, providing a richer interaction with the user through iterative feedback and refinement of results. During such iterations, the system can elicit the current user's preference and context. This in return has a positive impact on enabling users to better understand the search space, and reduce the effect of the cold start problem [11]. McCarthy et al. described CRS as a smart sales assistant, in which the recommender system is seen to play the role of a sales assistant who makes good suggestions. The sales assistant listens to the customer feedback and provides better suggestions in the next batch. In the same way a human agent should be able to provide satisfactory results in a reasonable amount of time based on iterative feedback, the recommender system should as well [1].

CRS uses two different strategies to help the user in navigation, *navigation by asking* and *navigation by proposing*; each relies on different forms of feedback. In navigation by asking, the user is asked to provide feedback on a feature. For example, the user provides a *value* in the feedback such as “I want a 2.5GHz processor” when looking for a laptop. Conversely, in navigation by proposing the system proposes a set of items and asks the user to provide feedback on the recommended items. For example, the user may *critique* a feature, such as “show me more like this but cheaper”, or may provide a *preference*, such as “show more items similar to this”, or may provide a rating for a set of the recommended items, such as “show me more items similar to this given that item X is 20% correct, and item Y 80% correct” [13]. Smyth and McGinty compared four forms of feedback used in CRS: Value Elicitation, Critiquing, Ratings-Based, and Preference-Based feedback. In *Value Elicitation*, the user specifies a specific value for a specific feature. In *Critiquing*, the user selects a directional preference for a certain feature in one of the recommended items. *Ratings-Based* feedback is the most common type of feedback, in which the user provides a rating for a subset of the recommended items. And finally, in *Preference-Based* the user selects one of the recommended items as the preference. Table 2.1, shows a comparison between the four types of user feedback in CRS. The cost to the user reflects the amount of thinking the user has to put in providing feedback. For example, in a car recommender system, asking the user to provide the precise number of fuel

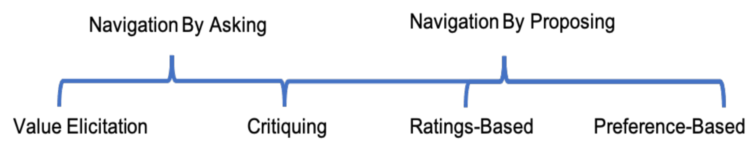


Figure 2.1: Types of CRS along with feedback type under each type

consumption per mile involves a higher amount of thinking, compared to providing feedback in the form of low, medium, or high fuel consumption. Therefore, the table shows a high cost for the Value Elicitation feedback due to the thinking process involved compared to Critique, Rating, or Preference feedback. Domain expertise is another factor that is related to the recommendation context. It refers to a users' ability to provide feedback depending on whether they have a clear understanding of the domain and what they are being asked for. In other words, the user should have a clear understanding of the domain to be able to answer a specific question about a specific feature in which the answer satisfies the user's needs. For example, in our example of car recommender, the user must know the meaning of fuel consumption and its relation to other car features such as speed, and motor size to provide a correct number that satisfies their needs. The main idea here is that users should have a minimum degree of domain knowledge to use a specific form of feedback. However, as shown in the table, the correct form of feedback should be carefully chosen. For example, Critique requires more domain knowledge from the user compared to Preference, and Value Elicitation requires more domain knowledge compared to Critique feedback. Figure 2.1, shows different forms of feedback along with its conversational type. Critiquing feedback has several variations of feedback forms and it falls under both types of conversational feedback, proposing and asking.

Table 2.1: A comparison between the four types of feedback regarding user's cost and expertise [1]

Feedback	Cost	Expertise
Value Elicitation	High	High
Critique	Moderate	Moderate
Rating	Moderate	Moderate
Preference	Low	Low

This dissertation uses critiquing forms of feedback. In critiquing feedback the user provides a directional preference over a feature of recommendation [2] [1]. For example, in a car recommender system, the user might ask for a smaller car engine than the currently recommended car. In this case, the smaller car engine is the critique over the car engine size feature. The form of critique also has potential variations: natural language dialog, user-initiated critiquing, or system-suggested critiquing [2]. The natural language dialog engages users in conversational dialog in the form of question-answer to capture preference feedback such as the recommender system proposed in [17]. In user-initiated critiquing, the system shows items and motivates the user to select features to critique such as the recommender system proposed in [18]. Finally, in the system suggested critiquing form, as its name suggests, the system generates critique over a set of features [19]. Since the goal is to increase diversity through the exploration process through critiquing, this dissertation adopts the system suggested critiquing form. Figure 2.2, shows an overview of the interaction between a user and a critiquing-based recommender system. In Step 1, the user is asked to specify the preference such as product features, reference products, or criteria. In Step 2, the recommender displays a number of items that satisfy the original preference. In Step 3a, the user at this point can stop the process and selects the recommended item which is unlikely to happen. In most cases, the user interacts with the system and provides a critique for some features as shown in Step 3b. Finally, in Step 4, the system applies the selected critique and recommends a new set of items closer to the user's preference.

Each critiquing form has its advantages and disadvantages, and this research focuses on system-suggested critiquing. The main advantage of this form of critique is its ability to educate and guide the user about the available items. This in return results in increasing the recommendation power. However, the accuracy measure of system suggested critiquing may be somewhat lower than preference-based recommenders. This is because the generated critique may not cover every possible scenario, and therefore users may require more iterations to find what they are looking for [20]. Burke et al. [19] developed seminal work on FindMe, a CRS with system-suggested critique. FindMe helps users to find products through a large multi-dimensional information space. The critiquing in FindMe proposed two challenges. First, it employs a pre-designed set of critique and fixed within the user interaction session, a form of critique called *static critique*. Second, each critique can only constrain one feature, so called *unit critique*. To overcome the first challenge, Reilly et al. have shown that standard critique can be extended to cover multiple features in what is called compound critique [21]. For the second-mentioned challenge, McCarthy et al. have developed a dynamic critique approach [22], in which the system combines features depending on the available items in the search space.

This dissertation proposes a novel approach to generate dynamic unit critiques in which each recommended item is presented along with a unique set of critiques that is different from the critique set of other recommended items. This work utilizes the interaction model presented by Chen in [2] as shown in Figure 2.2.

2.2 Diversity in Recommender Systems

The concept of diversity and its relationship to accuracy have been studied in Information Retrieval [23] and economics [24] before receiving attention in recommender

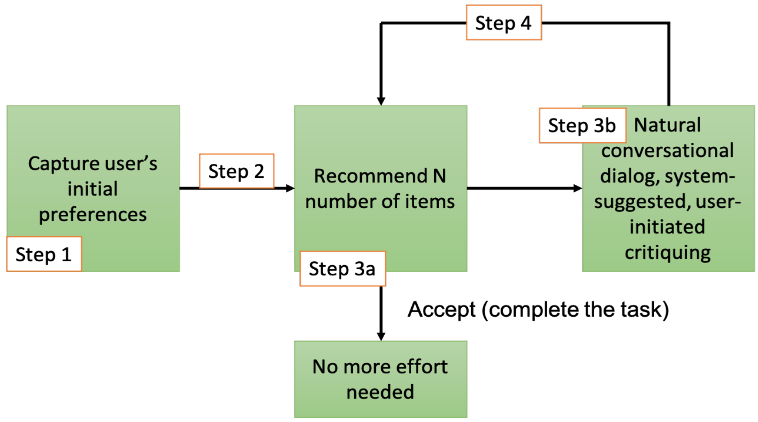


Figure 2.2: Interaction steps between user and critiquing-based recommender system [2]

systems. The idea of diversity introduced by Markowitz in Modern Portfolio Theory models investment as a tradeoff between risk and expected returns [24]. Maximizing the expected return results in higher risk, while diversification of stock portfolios reduces the risk. More generally, recommender systems adopted this idea where ranking items by their predicted relevance increases the risk of producing results that do not satisfy the user because the retrieved items tend to be too similar. On the other hand, diversifying the results reduces risk by increasing the probability of introducing items the user will be interested in [23]. Understanding the balance between accuracy and diversity is an important research question in recommender systems, with the emphasis that users are more likely to be satisfied with diversified results even with a small sacrifice in terms of accuracy [25].

2.2.1 Defining and Measuring Diversity

In Informtion Retrieval, the relevance of a retrieved document is not only affected by its similarity to the user's initiated query but with the similarity to other retrieved documents [23]. In particular, the role of diversity was used to solve the problem of ambiguity. For example, the word apple could refer to the company rather than the fruit. Therefore, retrieving a diverse set of documents is a way to solve ambiguous cases. In Recommender Systems, Bradley and Smyth [26] expanded upon this, defining diversity as the complement of similarity. Using this definition, Fleder and Hosanagar conducted a follow-up study and showed that recommender systems reduce diversity by focusing on accuracy [27]. Later on, Smyth and McClave suggested measuring the diversity of the list as the average pairwise distance [28]:

$$Diversity(R) = \frac{\sum_{i \epsilon R} \sum_{j \epsilon R/\{i\}} dist(i,j)}{|R|(|R|-1)} \tag{2.1}$$

where R is the recommended list of items, and $dist(i,j)$ is the distance between item i and item j.

Ziegler et al. [29] defined the Intra-List Similarity (ILS) metric using the aggregate pairwise similarity rather than the average, since similarity is used rather than distance; higher values of ILS indicate lower diversity. Using the pairwise distances between items to measure diversity has been widely adopted in the recommender systems literature with variations in the distance metric. These differences in distance metrics depend on an item's representation. For example, when items are represented by their content, the distance has been measured using the complement of Jaccard similarity [30], the complement of cosine similarity [31], or taxonomy-based metric [29]. When items are represented by rating vectors, hamming distance [32], the complement of Pearson correlation [30], or the complement cosine similarity have been adopted as a distance measure [33].

The average pairwise distance diversity metric in equation 2.1 was criticized by Vergas et al. [34], who argued that user's perception of diversity may not match the diversity produced by similarity metric. This argument is only valid in cases where items can be categorized into a set of genres such as books, or movies. Vergas et al. suggested using the genres to define diversity in such cases, arguing that diversity in genre aligns with the user's perception of diversity. To accomplish this, they proposed a genre-based diversity metric that considers three criteria, coverage, redundancy, and size awareness. In particular, a genre-based diversity score should reflect how well the recommended list covered the genre and avoided redundancy with regards to the recommended list size. The idea of considering the proportionality of the list size in diversity metric was adopted before by Dang and Croft [35], who considered the most diverse list of items is the list that the number of items that cover each topic is proportional to the topic's distribution in the search space.

2.2.2 Diversity in Recommender Systems

Two major techniques have been employed to incorporate diversity in recommender approaches. The first diversification technique involves reranking results generated by a recommender system. The second diversification technique involves the design of diversity-oriented algorithms. This section focuses on the reranking technique because of its relatedness to CRS.

The reranking diversification approach consists of recommending items R of size N from a candidate list of items C where $|C| > |N|$. The candidate items C are generated by an existing recommender algorithm to ensure the items' relevance. To construct the list R, most reranking approaches apply a greedy reranking algorithm, in which at each iteration the item in C maximizes an objective function to the list R. Algorithm 1 shows the general steps for a greedy reranking algorithm.

Algorithm 1: The Greedy Reranking Algorithm [14]

Input: N; a set of candidate items C s.t. $|C| < N$

Result: Result list R, s.t. $|R| = N$

R $\leftarrow$ [];

while $/R/ < N$ **do**

    i $\leftarrow \underset{i \epsilon C}{\text{argmax}} f_{obj}(i, R)$

    R $\leftarrow R + +[i]$

    C $\leftarrow C/\{i\}$;

end

return R;

The greedy reranking diversification technique has been adopted in studies such as [28, 29, 32]. In these studies, the objective function is defined as the linear combination of the item's relevance and diversity as shown in Equation 2.2:

$$f_{obj}(i, R) = \alpha.rel(i) + (1 - \alpha).\frac{1}{|R|}\sum_{j \epsilon R} dist(i, j) \tag{2.2}$$

where $rel(i)$ is the item's relevance, α parameter used to control the trade-off between relevance and diversity. The distance $dist(i, j)$ between item i and j can be computed using different approaches as discussed before.

In the context of case-based recommendation, Smyth and McCalve [28] applied the objective function in Equation 2.2 to retrieve items. The items were retrieved using the user's query. In that case, $rel(i)$ represents the item's relevance to the user's query, and $dist(i, i)$ represents the complement of the similarity between cases. Ziegler et al. [29], who adopted the idea of genre-based diversity, used the reranking technique for book recommenders. The relevant items were generated using the user's rating profile i.e Collaborative Filtering (CF). They defined $rel(i)$ as the book's relevance predicted by the recommender algorithm, and $dist(i, j)$ is the distance between books using a genre-based taxonomy.

Similar to other recommendation approaches, diversity was applied in conversational recommender systems as well. For example, Kelly and Bridge [32] applied a greedy reranking strategy in a conversational recommender system, the authors diversified the results in each iteration cycle after the feedback was received from the user. The item's relevance $rel(i)$ was generated from a CF recommender, and $dist(i, j)$ computed as the hamming distance of the two items' rating vectors. In another study, McGinty and Smyth [36] incorporated diversity in the conversational recommender system while balancing the tradeoff between diversity and relevance. The authors described a system where at each cycle, the user selects a critique which is used for the next iteration cycle. The selected item carried over the next recommendation cycle and was displayed along with other recommended items. If the user selects the carried-over item again the system assumes that no progress has been made and a more diversified list is recommended on the next cycle. However, if the user selects a different item, then the system assumes positive progress has been made and generates results with less diversity and more relevance for the next cycle.

As shown previously, one of the key advantages in reranking algorithms is the ability to separate the relevance part from the diversity part. It also allows for control of the trade-off between relevance and diversity as shown in Equation 2.2. However, incorporating diversity in CRS is a challenging problem because of the user feedback at each iteration. This dissertation proposes a new approach that combines dynamic critiquing with diversity in CRS. In comparison to previous studies, this research investigates an approach to generate critiques dynamically by selecting critiques that maximize diversity. The recommendations are generated depending on the user profile and preference, while the critiquing options are generated from diversity scores. Diversity scores are computed using established diversity metrics such as intra-list similarity. In this way, the recommender system will guide the user through the information space to more diverse items by selecting critiques that result in greater diversity.

2.3 Food Computing

Food computing is an interdisciplinary field that applies computational approaches to acquire, retrieve, and analyze data from different sources for perception, recognition, retrieval, recommendation, and monitoring of food to address food-related issues [37]. Figure 2.3 shows a general framework of the food computing field. Data can be captured from different sources such as social networks, cookbooks, and recipe-sharing websites. Collected data is analyzed using different approaches such as data mining, machine learning, computer vision, etc. The outcome of the analysis enables food-related tasks such as perception, recommendation, and retrieval, which can be

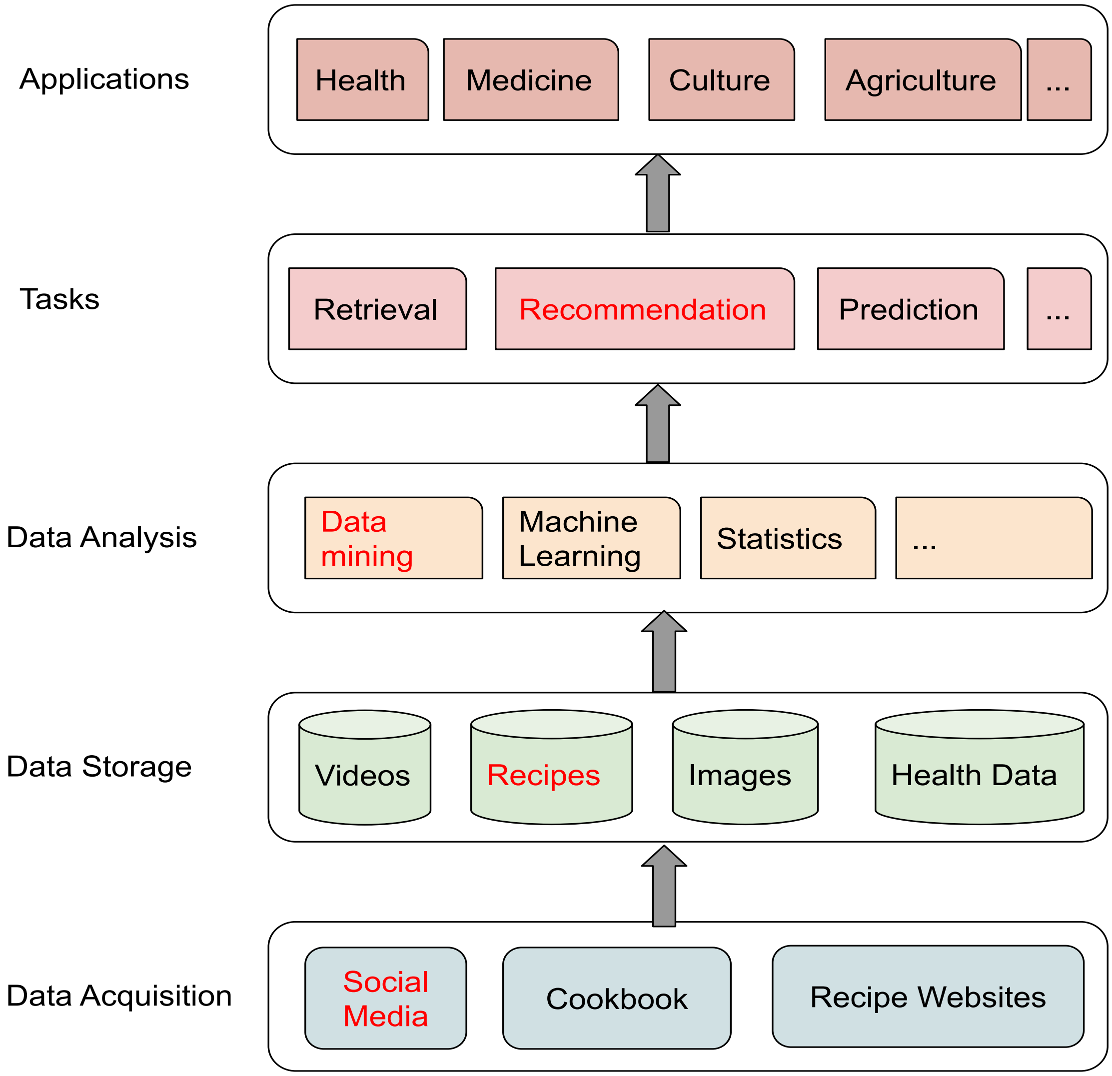


Figure 2.3: Food computing hierarchy [3]

used in different application domains such as health, agriculture, and biology. In this research, we are interested in recipe datasets collected from recipe-sharing websites, to perform recommendation tasks by through data analysis. The resulting recipe recommender engine can be used in the recipe search application domain.

2.3.1 Recipe Recommender Systems

Recommender systems have been proven to be useful in several application domains to help users in overcoming the burden of information overload, and assessing with the decision-making process [38]. One domain that has received comparatively little attention compared to other areas is the recommendation of food and recipes in particular. According to a survey conducted by Tratter and Elsweiler [39], there are five main methods found in the recommender systems literature that have been used to perform recipe recommendation tasks: Content-based methods, Collaborative Filtering-Based methods, Hybrid methods, Context-Aware methods, and Health-Aware methods. The next five sections provide an overview of the common recipe recommender systems found in each method.

Content-based methods Freyne and Berkovsky made recommendations by representing recipes as ingredients and providing scoring for each recipe depending on the user's ratings for recipes [40]. For example, if the user liked a recipe with tomato, the recommender recommends more recipes containing tomato. Harvey et al. [41] built upon this approach by negatively weighting recipes that contain ingredients that the user disliked. Other studies have adopted this approach by using different recipe representations such as topic model representation using recipe titles [42], and structure based representations using recipe instructions [43].

Collaborative Filtering-Based Methods (CBF) Freyne and Berkovsky used the nearest neighbor approach with Pearson correlation on the recipe-rating matrix to provide recommendations [9]. Harvey et al. [44] showed that using Singular Value Decomposition (SVD) provided better performance in comparison to the Freyne and Berkovsky approaches [9, 40]. Ge et al. [45] used a matrix factorization approach by leveraging users' ratings and tags to find recipes a user may like. They achieved better performance compared to content-based methods and standard matrix factorization approaches. Trattner and Elsweiler [46] ran a comparative studies between different collaborative filtering algorithms and showed that Latent Dirichlet Allocation(LDA) [47] and Weighted Matrix Factorization (WRMF) performed best for recipe recommendation [48].

Hybrid Methods Freyne and Berkovsky [9] combined a collaborative-based filtering method with a content-based method. In the study, they applied a hybrid approach that combines three different recommendation methods in a single model by using a switching strategy between the methods. Harvey et al. [41] achieved a better accuracy performance by combining their SVD approach with user and item biases.

Context-Aware Methods Food recommendation and exploration are highly dependable on contextual features. A number of studies in food and recipe exploration have focused on contextual features such as gender [49], time [50], hobbies [51], location [52], and food availability [53]. For example, Cheng et al. [52] explored location contextual factors by filtering users along with the location. In another study, Ahn et. al [54] studied the factor of culinary cultures in food recommenders. In their work, they have developed a flavor network to discover the pattern of ingredient combinations. Even though several studies addressed food recommendation from the contextual factors perspective, it remains an open question to understand which contextual factors are the most important factors in the food selection process [55].

Health-Aware Methods Health is always coupled with food recommendation research. Health problems and improving nutritional habits are usually mentioned with recipe recommendations [9, 40]. For example, Ge et al. [56] incorporated nutritional facts into a recommendation by accounting for calories into the recommendation. They achieved this by balancing between the calories the user needs and the calories the recipe has. In another study, Elsweiler et al. [57] developed an algorithm that balances the trade-off between what the user wants and what is nutritionally appropriate. Other studies incorporate popular healthy scores developed by *WHO* or *FSA* in the recommender algorithms [39].

Incorporating diversity into the food recommender system is a natural extension of Health-Aware recommenders. For example, Grace et al. [8] 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 focuses 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. In [57], Elsweiler et al. acknowledged the importance of diversity in meal plans as a way to provide healthy alternatives. They proposed a meal planner algorithm by recommending recipes to individuals. Despite the acknowledgment, they have not engineered diversity into the recommendations. However, the importance of diversity in recipe recommendation has several advantages such as: providing meals with varied sources of nutrition for a balanced meal diet [57], increasing user awareness of existing recipes [58], and covering a wide variety of options that could reduce the cold start problem [37].

2.3.2 Recipe Datasets in Food Computing

Many popular recipe datasets have been developed and released. Table 2.2 lists the most common datasets found in literature. The size column shows the number of recipes in the dataset, the source column lists the dataset source, and the data column shows the data found in each dataset. Most of the recipes were crawled from recipe sharing websites such as yummly.com and allrecipes.com. The majority of datasets consist of tens of thousands of recipes. At the time of writing, Recipe1M is the largest released recipe dataset with 1 million recipes and 800K images. The table also shows a variety of the available meta-data for each dataset such as videos, images, ingredients, regions, and meal courses. This dissertation uses the Recipes157K dataset because it incorporates many different types of features such as flavor, nutritional info, ingredients, cuisine, and meal course. More details regarding the selected dataset will be covered later in Section 4.2.

Table 2.2: A comparison between available recipe datasets

Dataset Name	Size	Source	Data
Recipes56K [54]	56,498	epicurious, allrecipes, menupan	Ingredients and Cuisine
AllRecipes46K [59]	46,33	allrecipes	Ingredient, Region, Nutrition, and course
RecipeSource5K [60]	5917	recipesource	Ingredient, Region
VireoFood172[61]	119241	Go Cooking, Meishi	Ingredient, Images
Recipes157K[62]	157 K	Yummly	Ingredients, Cuisine, Course meal, Nutritional Info, and flavor
Go Cooking[63]	61139	xiachufang.com	Images
Recipe1M[64]	1 Million	20 Cooking Website	Image, Ingredients, Directions
Yummly-28K[65]	28K	Yummly	Image, Ingredients, Cuisine, and Course meal
Yummly-66K[66]	66K	Yummly	Image, Ingredients, Cuisine, and Course meal
Recipes242K[67]	242,113	Crowdsourcing	Image, Ingredients, and Directions

2.4 Summary

Introducing diversity into recipe recommenders has been identified as an important research problem, which has received comparatively little attention. CRS enables the user to guide the direction of recommendation, and this dissertation focuses on investigating the CRS approach to recommendation as part of diet diversification, which can help to address several challenges in the recipe recommendation process.

CHAPTER 3: Generating Dynamic Critique to Support Diversity

Critique generation process involves finding a set of critique options for the user to select from. The selected critique serves as a feedback that the recommender systems uses to refine the next iteration of recommendation. This chapter presents a critique generation approach that generates a set of critique options for each recommended item.

3.1 Overview

Recipes in recommender systems can be represented using different features such as user rating, or ingredients. This dissertation focuses on content representation rather than user ratings. A large set of critiques can be generated for a single recommended item. This set can hold all the features that are related to the item. Critiques can come in the form of comparison, with three main options for on the same feature: 'less than' (<), 'greater than' (>), or 'similar to' (==), and this increases the space of possible critiques that can be used for a single item. To limit the number of critiques, two major approaches are presented in the literature: static critique, and dynamic critique. In static critique, a finite set of critiques are pre-determined and displayed for every recommended item in every iteration. In dynamic critiquing, different critiques are generated during the recommendation process based on metrics such as frequency, and relevance [1]. This dissertation adopts a dynamic critiquing approach, and introduces a novel approach for both selecting and directing candidate critiques. The approach proposed in this chapter is designed in the context of recipe recommendation, but the approach can be applied to any application domain with a careful selection of representation and critique features.

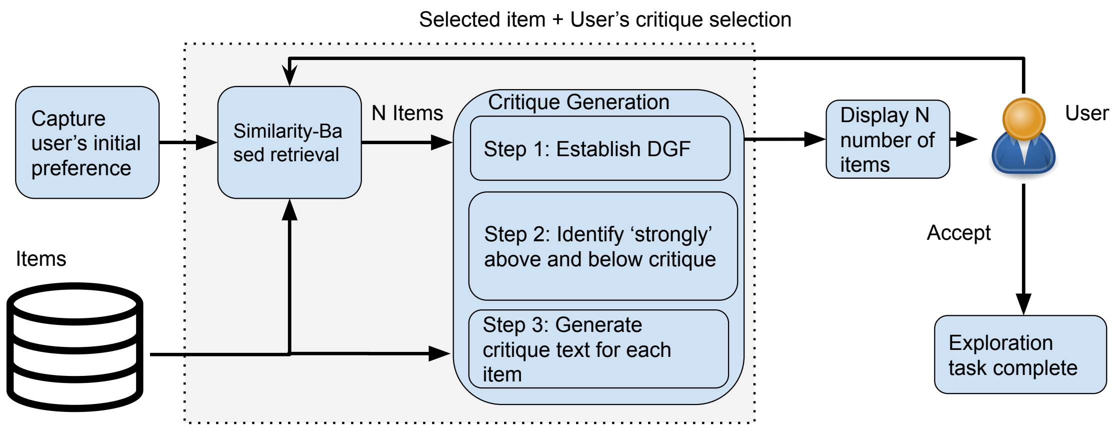


Figure 3.1: DiversityBite model, starts with user initial preference and ends with user acceptance. The shaded area represents retrieval and diversity cycles through critique.

The following section describes the proposed DiversityBite approach, a conversational recommender framework used to recommend diverse recipes, including the major components for baseline recommendation and diversity-focused critique generation.

3.2 DiversityBite: A Conversational Recipe Recommender System

The DiversityBite approach adopts an initial zooming stage [19] based on user preference, followed by a series of conversational interactions in the form of critique-based *recommend-review-revise* cycles [21]. This is enabled by two primary components: retrieval and critique which complement each other to address the similarity vs. diversity balance [28]. The DiversityBite framework is modular and can support a variety of representations for user profiles, as well as different metrics for similarity, retrieval, and critique [68, 69, 70].

Figure 3.1 shows the DiversityBite model. The interaction process starts by capturing the user's initial preference. In the recipes domain, the preference can be as general as cuisine type or specific as an excluded list of ingredients. The similarity-based retrieval retrieves similar recipes to the user profile. The critique generation step generates critique options for each retrieved recipe by following 3 main steps:

Zooming Stage, building a Diversity-Goal, and Critique Generation. The Diveristy-Bite framework is designed in a modular fashion in which each component model can be replaced. For example, similarity-based retrieval can be replaced by any content or collaborative-based filtering. This chapter focuses on describing each step in the following sections.

3.3 Similarity-Based Retrieval

Similarity-based retrieval considers the user's preference and recipe match. Initial user preference can be a hard constraint that limits the search space of available recipes such as cuisine type, or a soft constraint that penalizes the objective function if the condition is not being satisfied such as meal course. The degree of similarity is based on the recipe representation match such as ingredients. For a current recipe (r_c) and candidate recipe for retrieval (r_t):

$$sim(r_t, r_c) = \begin{cases} 0 & \text{if } r_t \text{ does not satisfythe hard constraints} \\ sim_{content}(r_t, r_c) + & \\ sim_{soft}(r_t, r_c) & \text{otherwise} \end{cases}$$

The $sim_{content}(r_t, r_c)$ metric is straightforward cosine similarity depending on the recipe representation. The $sim_{soft}(r_t)$ metric is the proportion of user-specified soft constraints found in the user profile that match the recipe r_t.

Recipes can be represented using a variety of different features such as ingredients, preparation steps, nutrition details, or user ratings. This dissertation focuses on a content-based representation. Recipes are represented with three distinct sets of features: ingredients, nutritional information, and flavor characterization. Section 4.2 provides more details on the recipe representation.

The output of the similarity-based retrieval model is a list of N recipes. The list

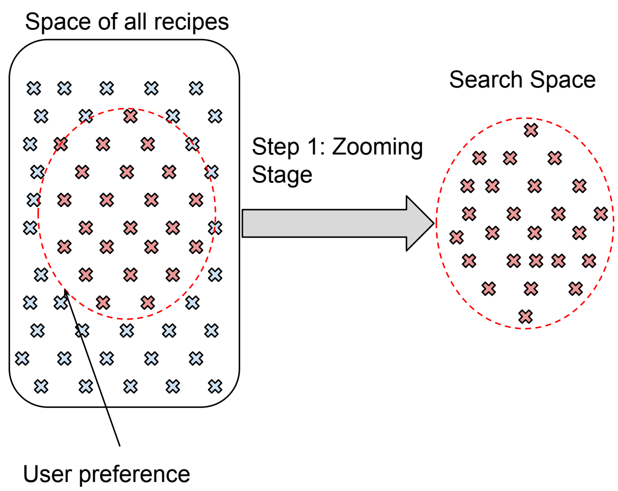


Figure 3.2: Limiting search space

of recipes serves as an input for the critique generation model discussed next.

3.4 Generating Dynamic Critique to Support Diversity

The DiversityBite approach to generate dynamic critique is shown in Figure 3.1, consisting of three main components: Zooming Stage, Building Diversity-Goal, and Critique Generation.

3.4.1 Step 1: Zooming Stage

The Zooming Stage is the first step to generate meaningful critiques through defining a search space. Defining the search space among the set of recipes eliminates outliers, and ensures baseline relevance of recipes. While limiting the search space may reduce some potential diversity, it ensures a foundational balance between diversity and preference. A very narrow search space will result in less diverse recommended recipes, while a wide search space may increase the potential for diversity in the recommended recipes. Figure 3.2, shows an imaginary red dotted line representing recipes within user preference. The result is a set of recipes that represent the user's recipe search space.

3.4.2 Step 2: Building Diversity-Goal

The second step is to understand the diversity among the recipes within the space. Once the user's initial preference has been determined, DiversityBite establishes a diversity goal (DG). This is a set of **S** recipes that (1) meet the user's baseline preference and (2) are selected for high diversity within the set. The DG approximates the maximal potential diversity among recipes within the current search space for the user's query. So, when a user selects a critique of a current recipe (essentially, "like this, but more diverse"), the DG provides a reference for selecting directions to move toward in the search space that is expected to increase diversity in recommendations. This dissertation investigates two variations of DG, a Diversity-Goal Footprint (DGF) and an Adaptive Diversity-Goal (ADG).

3.4.2.1 Diversity-Goal Footprint

The Diversity-Goal Footprint metric estimates maximum diversity by considering the whole search space. This variation does not consider the retrieved set of recipes in estimating the maximal potential diversity. Instead, it estimates the maximal potential of diversity by considering all recipes within the user's profile preference. DGF provides a faster computation time and an approximation to the most diverse list of recipes found in the search space.

3.4.2.2 Adaptive Diversity-Goal

The Adaptive Diversity-Goal metric estimates the maximum diversity by considering the whole search space. However, this variation considers the retrieved set of recipes as part of estimating the maximal potential diversity. It estimates the maximal potential of diversity for each retrieved recipe. Therefore, for each recommended recipe, a different set **S** is estimated dynamically, hence "Adaptive". In contrast to DGF, ADG requires more computational time, but the generated recipe set **S** considers the actual recipe and the search space as well.

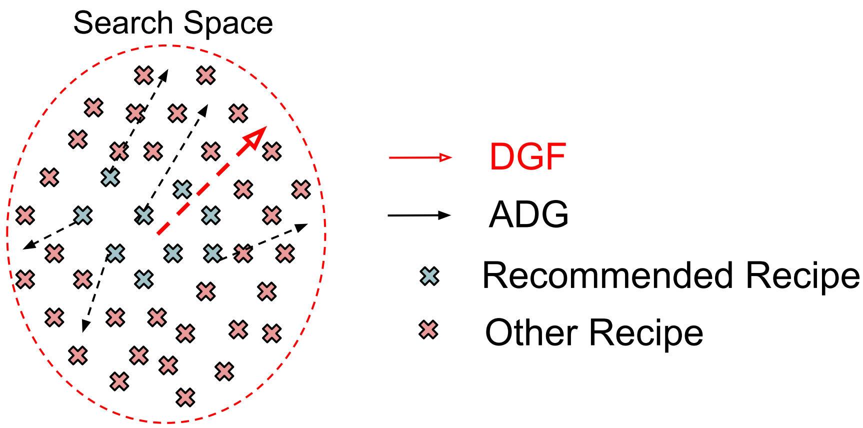


Figure 3.3: The difference between DGF, and ADG

Figure 3.3 shows a comparison between DGF and AGF. The vector represents an imaginary line toward a more diverse search space. The vector for DGF is computed once for the set of recommended recipes. While in the case of ADG each recipe has its own estimated DG.

3.4.3 Step 3: Generate critique

Given a list of recommended recipes from the Similarity-Based Retrieval model, a set of potential critiques is dynamically generated for each recipe. These are selected as possible pathways for the user toward greater recipe diversity in the next step of the conversation. Only critiques toward greater diversity are presented as actionable options for the user.

The potential critique dimensions correspond to features related to the recipe such as flavor and nutrition (e.g., more spicy or fewer calories). In both DGF and ADG approaches, the DG set **S** is used as a reference point for critique direction (more / less) across the critique dimensions. First, a threshold value of each critique dimension is taken as a baseline threshold. In the case of DGF, the threshold values were taken across all recipes in the search space. While in the ADG the threshold values were taken across the selected set **S**. Second, for each critique dimension, the percentage of recipes in the DG that are above or below the baseline threshold is recorded.

The **M** highest-percentage critique dimensions above and below the threshold are considered to be 'strongly' above and below the threshold for purposes of critique activation. This represents a directional vector used to guide critique toward greater diversity in results. Third, the current recipe's value on each critique dimension is checked in relation to the threshold value and the DG profile for that dimension. If the current recipe's value is below the threshold and the DG percentage for that dimension is strongly above the threshold, then a 'more' critique for that dimension is activated. Conversely, if the value is above the threshold and the DG percentage for that dimension is strongly below the threshold, then a 'less' critique for that dimension is activated.

Finally, DiversityBite displays a list of recipes along with its critique set. The user has the option to (1) make a decision regarding the displayed recipe, and (2) select a recipe + critique that will be used as feedback for the next round of conversation. An applied critique serves as a filter that is applied to similarity-based retrieval - filtering recipes from the top **N** that do not satisfy the critique with respect to the reference recipe.

3.4.4 Summary

This chapter presented DiversityBite, an interactive framework that dynamically generates critiques for a set of retrieved recipes to promote diversity during exploration. The framework consists of two major models along with a feedback loop from the user. The first model consists of a Similarity-Based Retrieval model that selects recipes that match the user's profile. The second model generates critique for each retrieved recipe for the user to select from. The critique generation model consists of three stages. The first stage is a zooming stage that depends on the user profile, the purpose of this stage is to set the boundaries of the exploration to be within the user's preference. The second stage is to estimate the maximal potential diversity within the current search space by defining a diversity goal (DG). The DG can be general and related to the search space regardless of the retrieved recipes (DGF) or tailored for each recipe and changes with every recommendation (ADG). The final stage is the critique generation, in which each critique dimension is identified to be above or below a specific threshold for critique activation. The next chapter presents the evaluation conducted on DiversityBite to measure its effectiveness in recommending diverse recipes.

CHAPTER 4: Evaluation

4.1 Overview

This chapter presents the evaluation methodologies used to evaluate DiversityBite. This dissertation adopts two primary lines of evaluation strategies, offline and online. The offline evaluation serves as a basis for more sophisticated evaluations by simulating real scenarios. The online evaluation involves recruiting real users to experiment with a fully implemented version of the DiversityBite framework where user actions and feedback are logged for further analysis and exploration.

This chapter starts by describing the dataset used in the experiments, followed by an overview of the evaluations where it presents the offline evaluation followed by the user studies. The purpose of the offline evaluation is to establish a basic understanding of the effect of incorporating diversity during the critique process. The user studies focus on evaluating the effect of dynamic critiquing on diversity as well as the effect of recipe representation on diversity.

4.2 Dataset Description

This work uses a recipe dataset that has the potential of diversity. In [62], Sajadmanesh et al. prepared a dataset with 120K recipes crawled from yummly.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 identified by six flavors, namely, saltiness, sourness, sweetness, bitterness, spiciness, and savoriness. The flavor features are represented on a scale from 0 to 1. Compared to other datasets, the variety of recipes from different countries in the dataset makes this dataset suitable to discover diverse recipes. This dataset is also rich in metadata such as flavor and nutrition that can support the critique generation approach proposed in Chapter 3.

Figure 4.1 shows the distribution of recipes over cuisines. The most common cuisines are Mexican, Italian, American, Indian, and Irish. The least common cuisines (not shown in the figure) are Benin, Eritrean, and Chadian. The average number of recipes per cuisine is 466 recipes with a high standard deviation of 941 meaning that the dataset has a long tail of cuisine with a small number of recipes. The purpose of the dissertation is to promote diet diversification while exploring recipes, hence all cuisines were considered in the evaluation. Figure 4.2 shows the range of possible values for each flavor feature. For each flavor, most recipes fall within the medium range except for Spiciness in which recipes tend to be less spicy. Saltiness has the highest variation among all flavors. Figure 4.3 and Figure 4.4 show the value distribution of nutritional facts found in the dataset (separated in two figures for clarity). The nutritional value here is shown per serving. Figure 4.5 shows a heatmap with Pearson's correlation for both nutritional and flavor features. For the nutritional features, the figure shows a strong positive Pearson's correlation between Fat, Saturated Fat, and Calories. The figure also shows a strong positive correlation between sugar and carbs. Among the flavor features, there is only one strong positive correlation between Bitterness, and Saltiness, while Saltiness is negatively correlated with Spiciness. The heatmap and the box plots do not show any inconsistency or erroneousness in the data. Therefore, all features are considered during the evaluation.

To reduce overall sparsity in ingredients, FOODON [71] ontology was used to map each ingredient to a food concept. The mapping reduced the number of unique ingre-

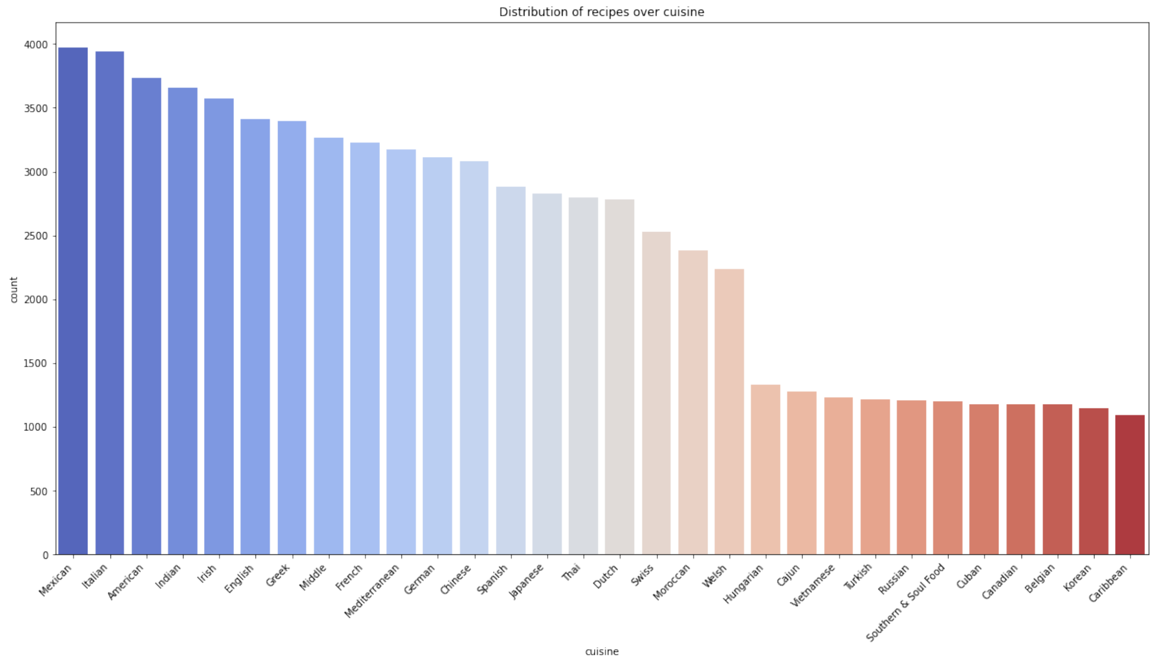


Figure 4.1: The most common 30 cuisines found in the dataset

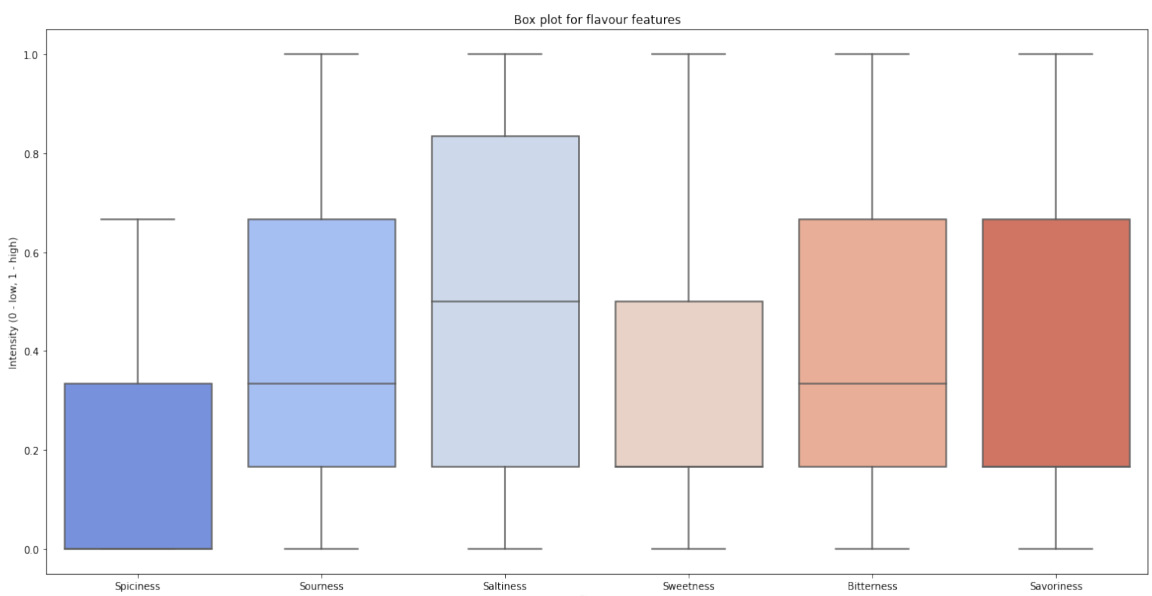


Figure 4.2: Boxplot shows the range of values for each flavor feature.

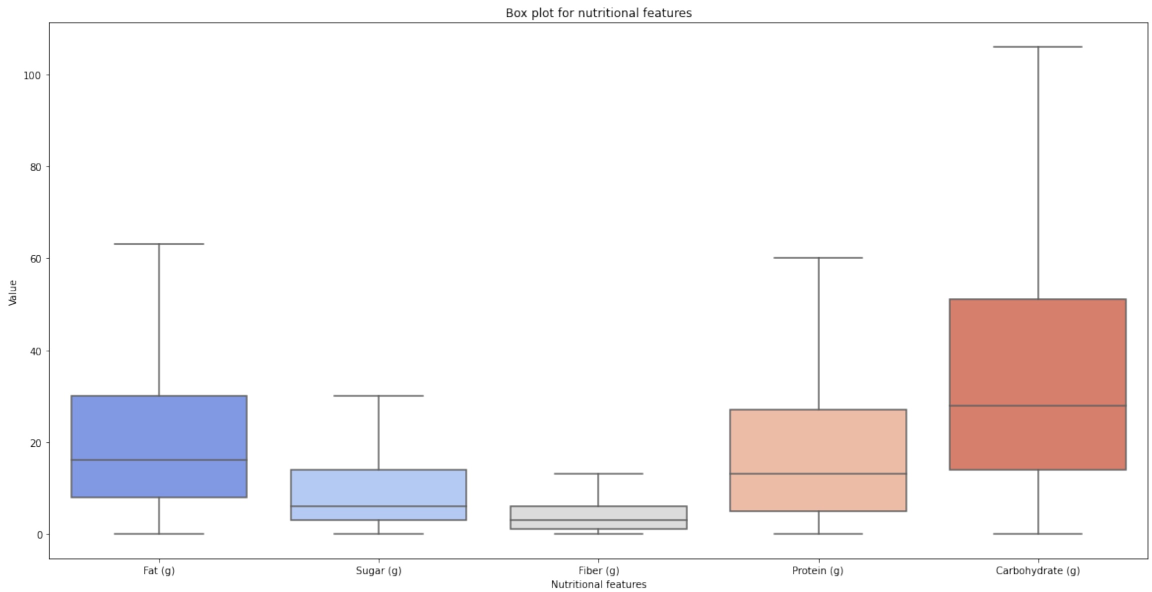


Figure 4.3: Boxplot for the nutritional features: Fat, Sugar, Fiber, Protein, and Carbohydrate

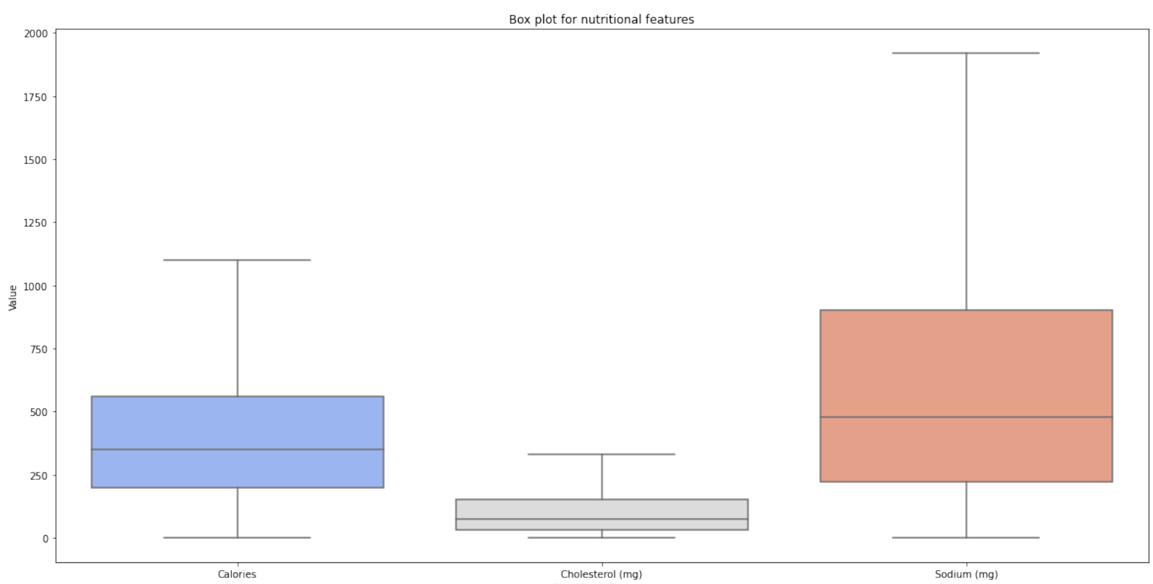


Figure 4.4: Boxplot for the nutritional features: Calories, Cholesterol, and Sodium

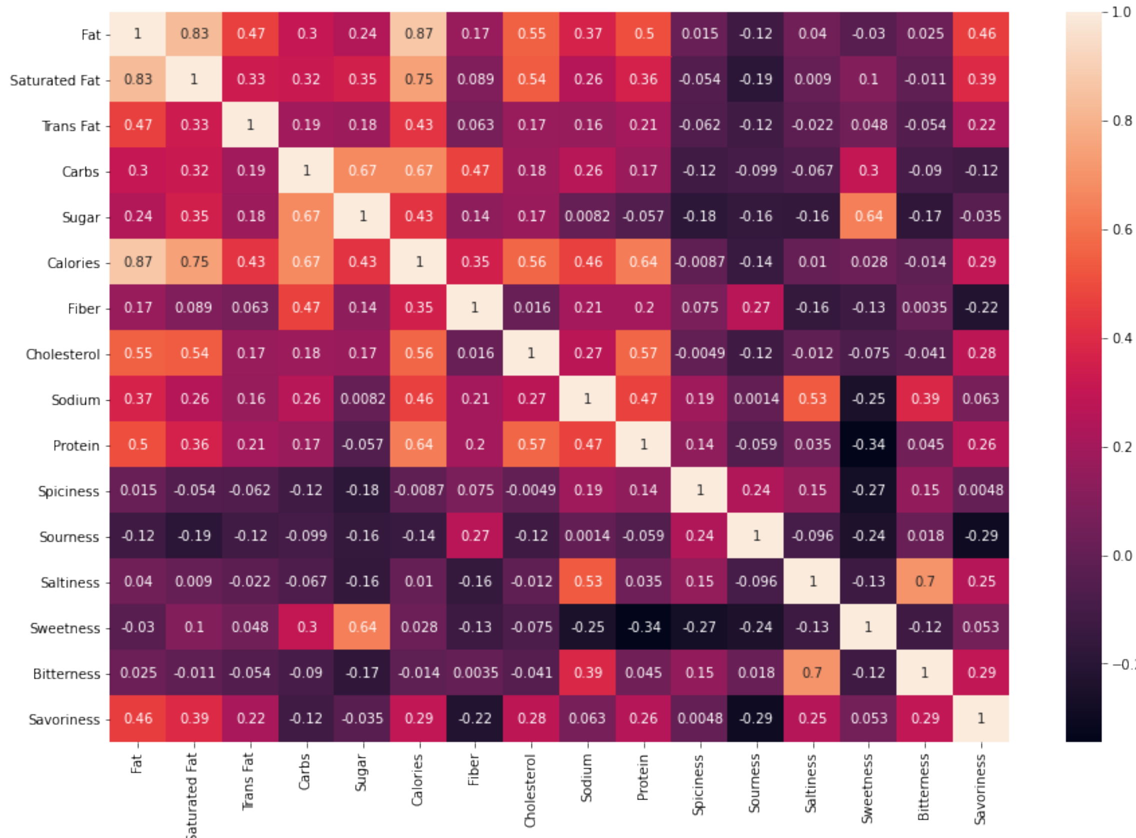


Figure 4.5: A heatmap shows Pearson correlation between all nutritional and flavor features

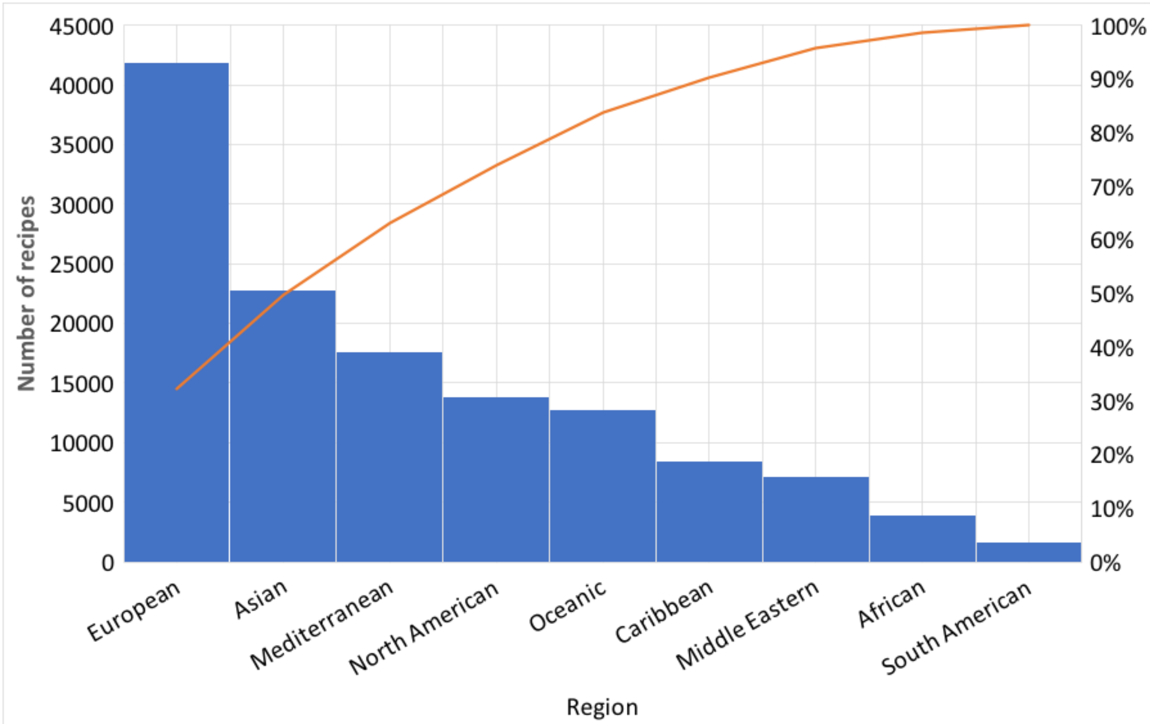


Figure 4.6: The distribution of recipes over the region in the dataset

dients from 11,113 ingredients to 3,807 ingredients. To facilitate user interaction and capture the user's initial interest we have grouped recipes by region into 9 regions: Caribbean, North America, South America, Europe, Africa, Middle East, Mediterranean, Asia, and Oceanic. The grouping of regions is based on mapping each country to a specific region by following the continental regional cuisine classification in [1]. For example, all Egyptian recipes are mapped to Africa region. Figure 4.6 shows the distribution of recipes over these regions.

4.3 Implementation Specification: DGF, AGD, and Diversity Scoring

There can be several implementations for the Diversity-Goal Footprint (DGF) and the Adaptive Diversity-Goal (ADG). The output of both approaches is still the same, which is a list of **S** recipes with high diversity. This subsection discusses the implementation for DGF and ADG used in the evaluation studies.

To estimate DG list using DGF, **L** number of recipes were randomly selected from

[1]https://en.wikipedia.org/wiki/Regional$_{c}uisine$

the search space and the diversity score was calculated for the generated list. This process was repeated **R** number of times. Then, the list with the highest diversity score was selected to represent the DG. This approach follows Vergas et al. [34], who note that maximum diversity can be approximated through random selection. In addition to that, random selection provides an advantage in terms of computation time.

In the case of ADG, the DG was estimated using the Greedy re-ranking algorithm. The implementation follows Dijkstra's shortest path algorithm [72] to select the next recipe to be added to the list **S**. In particular, given a current recipe **r** the next recipe **n** should be the farthest (most dissimilar) recipe from **r** within the search space. This approach requires the computation of distances between recipes in the search space and therefore requires more computational time. However, the heavy load of the computation can be completed offline to alleviate this problem. In this work, the K-D tree algorithm [73] was used to calculate the distances between recipes. K-D tree is a space partitioning data structure that separates the space into K dimensions. This data structure is very effective in nearest neighbor search applications by using tree properties to eliminate a large portion of the tree in the search process. Since the AGD approach requires the recommended recipes as input, each recipe in the recommended Top-N recipes serves as the start node of the shortest path algorithm to estimate DG.

The diversity score is calculated using the average pairwise distances between recipes following Smyth and McClave [28], as shown in equation 4.1.

$$Diversity(R) = \frac{\sum\limits_{i \epsilon R} \sum\limits_{j \epsilon R/\{i\}} dist(i,j)}{|R|(|R|-1)} \tag{4.1}$$

The details of $dist(i, j)$ measure are mentioned along with the study-specific implementation details in each of the experiments sections.

4.4 User Study Design

This section discusses the flow of the two user studies, both studies share a common flow with minor variations that will be highlighted as each experiment is presented. Figure 4.7, shows the general workflow in both studies. Participants were asked to fill out a pre-survey about their demographic information (age, gender, and education), and their online behavior while searching for recipes. After the initial survey, participants were asked to use the recommender system to prepare a week-long meal plan. The task prompt was: "For the next system variations, prepare a meal plan for a week". During the task, the application interface displays two progress bars as shown in Figure 4.9. The first progress bar indicates exploration progress and its maximum value achieved after 7 explorations, while the second progress bar indicates the meal plan completion and its maximum value achieved after adding 7 recipes to the meal plan. However, participants can explore and add more recipes to the meal plan but the progress bars ensure a minimum of 7 cycles, and 7 meals are added to the plan before ending the session and moving to the next variation. We note here that there are no restrictions on the meal plan requirements other than having at least 7 recipes. The recipes could be any kind of meal types such as main-dish, or salad. The interface also shows a search history for previous exploration results as shown in Figure 4.9. Figure 4.8, shows a screenshot for one of the DiversityBite variations displaying a list of recipes recommended to the user, as well as an example of the expanded view where the user can display additional information for a particular recipe. The user can explore more recipes by selecting one of the displayed critiques, they can also dislike a recipe so it will not appear in any upcoming iterations while exploring.

The studies followed a within-subject design, in which each participant had to repeat the same task as many times as the number of variations of DiversityBite. The view order of the recommender variations was counterbalanced to account for

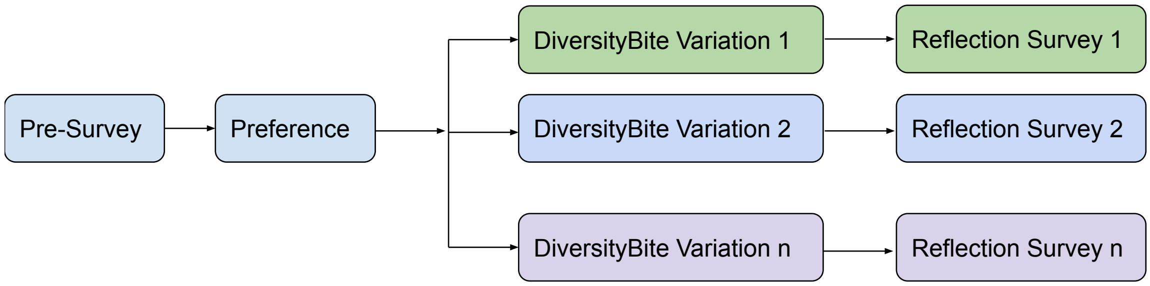


Figure 4.7: The flow for both user studies

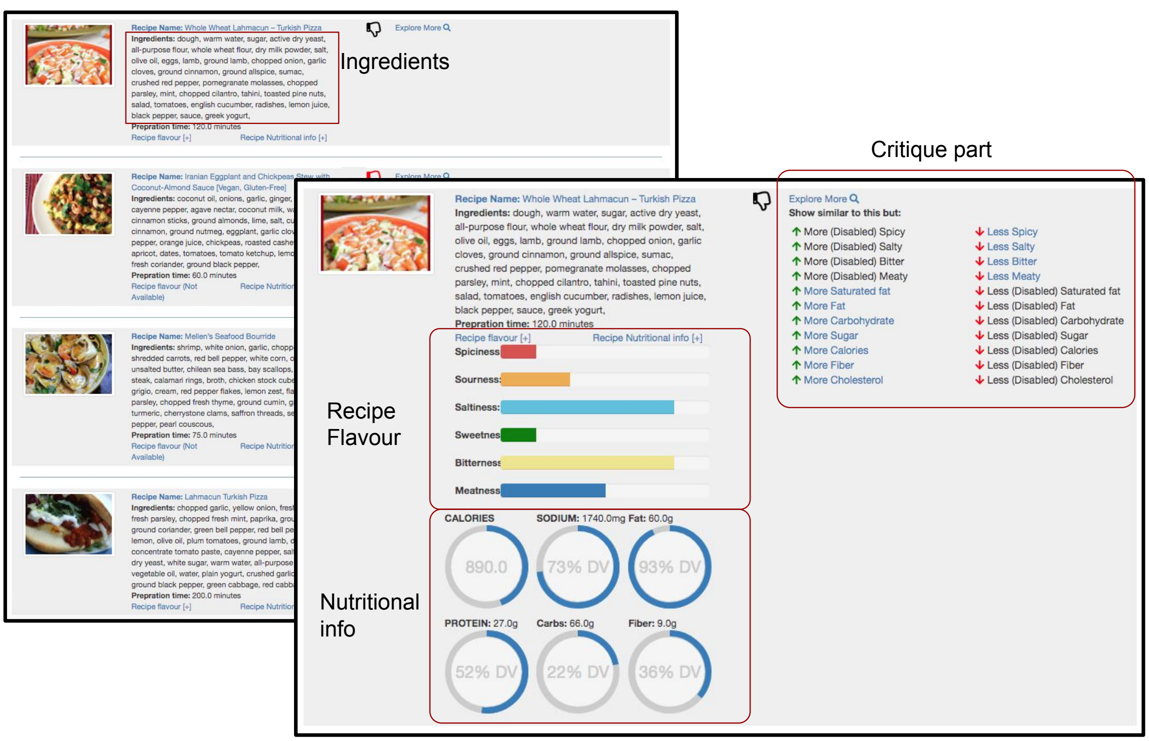


Figure 4.8: A screenshot of one of DiversityBite variations. It shows four recipes, the enlarged image shows the recipes details i.e. flavor features, nutritional features, and critiques. The critique part shows only critique that can lead to more diverse cases.

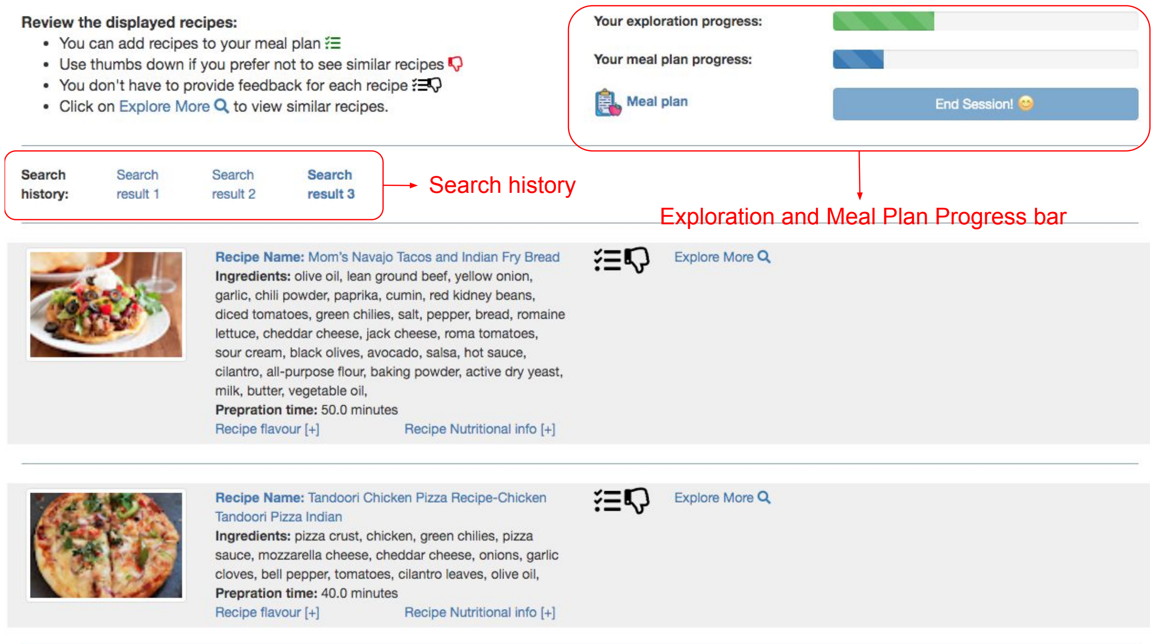


Figure 4.9: Screenshot of DiversityBite interface shows search history and the progress bar for exploration progress and meal plan progress.

participants' fatigue and learning effects. To indicate their preference, participants were asked to select one or more regions/cuisine they would like to see recipes from, and indicate their meal course preference. The same preference selection was used in all recommendation variations. After each recommendation variation, participants were asked to fill out a reflection survey based on their experience.

4.5 Simulation Study: Incorporating Diversity in Critique

4.5.1 Overview

To address the first research question, a simulation study for critique generation was conducted. To recall the research question is:

RQ 1: How do critique-based approaches impact the diversity of recommendations?

- **H 1.1** Critique-based recommenders result in recommending more diverse recipes compared to non critique-based recommenders.

- **H 1.2** Critique-based recommenders achieve higher diversity scores in less number of iterations compared to non critique-based recommenders.

The purpose of the simulation study is (1) to understand the diversity scores of the recommended recipes over the course of the users' interaction, 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 the evaluation results.

4.5.2 Experiment Setup

To evaluate the critique approach, three variations of DiveristyBite were built. One variation without critique (DiversityBite-) simulates a similarity-based recipe recommender, the other two variations incorporate critique. The first critique variation implements the GDF (DiversityBite+GDF), while the other critique variation implements the ADG (DiversityBite+ADG). The algorithm to recommend recipes in all variations was kept consistent; the only difference is the presence of critique. The first iteration in the recommendation starts recommending the closest N recipes to the centroid of the user's 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 is represented by a vector of 3,807 ingredients. The ingredient vector is a binary vector with 1 means the ingredient is present and 0 otherwise. In the case of DiversityBite+GDF and DiversityBite+ADG, the proposed algorithm to generate critique was used. The GDF approach was used in DiversityBite+GDF variation, while ADG was used in DiversityBite+ADG variation. After the selection of a recipe and a critique, the algorithm recommends N closest recipes to the selected recipe with the critique applied.

To comprehensively evaluate the impact of diversity in the recommendation on users, this work recognizes that user study evaluation will be needed. As a first step in that direction, this section discusses an offline simulation to serve as a baseline for a more comprehensive user study. The simulation consists of building 100 user profiles. Each profile is evaluated by simulating 50 iterations using DiversityBite-, DiversityBite+DGF, and DiversityBite+ADG. Given that the yummly.com dataset, as discussed before, does not provide user interaction with recipes, user profiles were created by randomly selecting a region that consists several cuisines. Then, a random fraction of recipes with an average rating of 4 or more (on a typical 1 - 5 scale) were selected to build the user profile. The search space for each user is the rest of the 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 terms of critique selection, due to the lack of data of users' preference of critique a random critique was chosen among the critique list of the selected recipe. Selecting a critique randomly among the set of the generated critiques adds randomness to the process. Therefore, there is no guarantee during the simulation that the selected critiques will fully match user's preference. To reduce this effect the number of iterations for the simulation was set to 50. While a high number of iterations doesn't reflect a real scenario it gives the opportunity to study a prolonged interaction. It also minimizes the effect of choosing critique randomly as the randomness of choosing a critique will follow a normal distribution based on the central limit theorem.

For the implementation of the three variations the following settings were used: $\mathbf{N} = 10$, $\mathbf{K} = 3$, $\mathbf{L} = 100$, $\mathbf{R} = 100$. The total number of unique critiques is 16 (6 flavors + 10 nutrition). The threshold value $\boldsymbol{\varepsilon}_{\mathbf{C_n}}$ for each critique was selected to be the average value of the feature that represents the critique. 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. Recipes were represented as a binary vector of ingredients with value of 1 means the ingredient is present in the recipe.

Cosine similarity was used to calculate similarity between recipes. To estimate the diversity goal (DG) in DiversityBite+ADG variation the K-D tree data structure uses a euclidean distance to find the nearest neighbour.

4.5.3 Analysis and Results

This section reports on findings related to the first research question by addressing the related hypotheses. The first section addresses the first hypothesis (H 1.1) while the second addresses the second hypothesis (H 1.2).

4.5.3.1 Diversity Improvement Analysis

After running the simulation, the diversity score for the recommended set of recipes was measured at each iteration using the diversity equation 2.1. Figure 4.11 shows the diversity score of the first 15 iterations for the same user in DiversityBite-, DiversityBite+DGF, and DiversityBite+ADG. As shown in the figure, in all iterations (except one iteration) the diversity score from DiversityBite+ADG is always higher than the diversity score for DiversityBite-, DiversityBite+DGF. The reason behind having lower diversity scores sometimes is that the simulator chooses a critique with a lower diversity score at one iteration compared to the rest. However, the overall diversity score for other iterations shows that DiversityBite+ADG is higher. Figure 4.10 shows the overall distribution for each variation. To address the first hypothesis (H 1.1), 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 scores in DiversityBite- ($M=2.08$, $SD=0.39$). These results show that using critique enables the recommendation of more diverse recipes. However, the results here show no statistical significance between DiversityBite+DGF, and DiversityBite+ADG suggesting that both methods can be used to

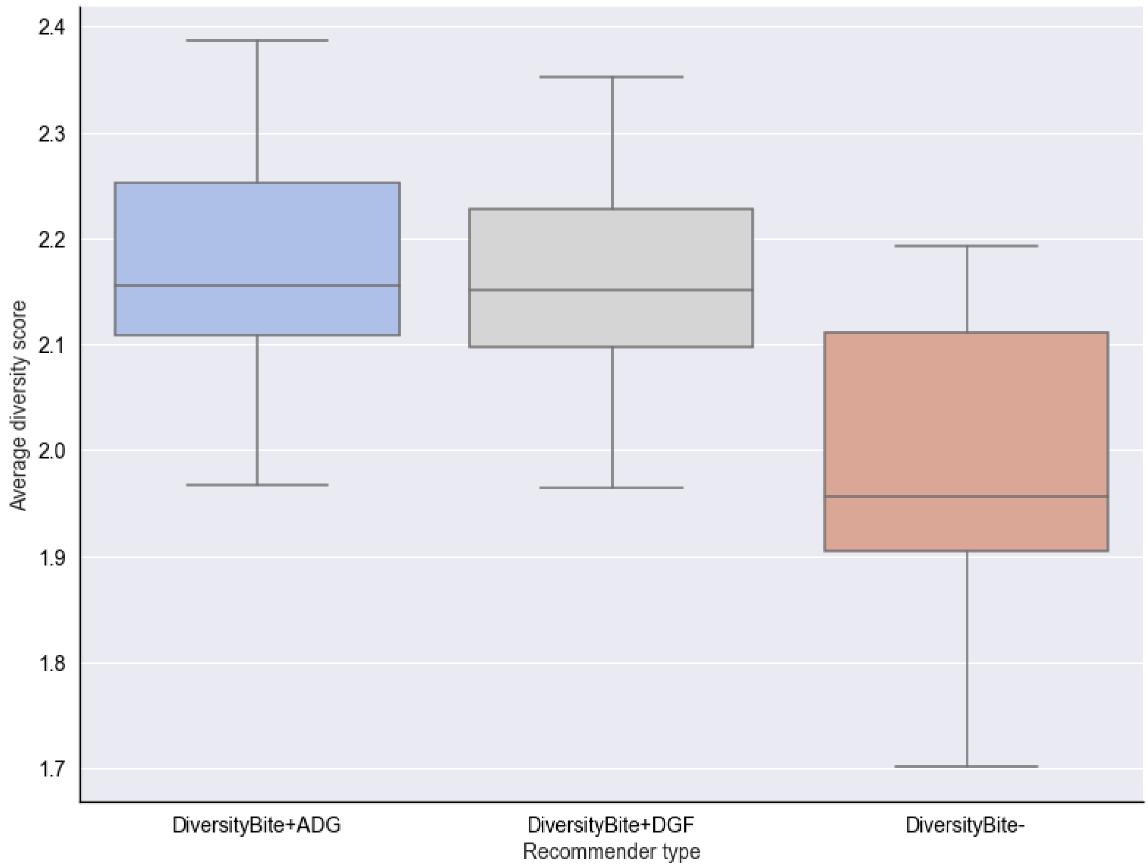


Figure 4.10: A comparison between DiversityBite-, DiversityBite+DGF, and DiversityBite+ADG diversity scores

increase diversity compared to the baseline DiversityBite-.

4.5.3.2 Diversity Improvement and Number of Iterations

In the previous section we examined the diversity scores of recommended recipes under different variations of DiversityBite. To address the second hypothesis H 1.2, this section studies the relation between the number of iterations and the diversity scores. The results show that starting at the second iteration there is a statistical difference between the critique-based recommender (DiversityBite+DGF, and DiversityBite+ADG) and the non-critique recommender (DiversityBite-). Figure 4.11 shows that the p-value is less than 0.05 after the second iteration. The results at each iteration confirm with the findings in Table 4.1 in terms of post-hoc analysis in which the

Table 4.1: Detailed results of Tukey's post hoc test shows that diversity in DiversityBite+DGF, and DiversityBite+ADG is significantly different than the diversity scores in DiversityBite-

Variation 1	Variation 2	p <0.05
DiversityBite+DGF	DiversityBite+ADG	False
DiversityBite+ADG	DiversityBite-	True
DiversityBite+DGF	DiversityBite-	True

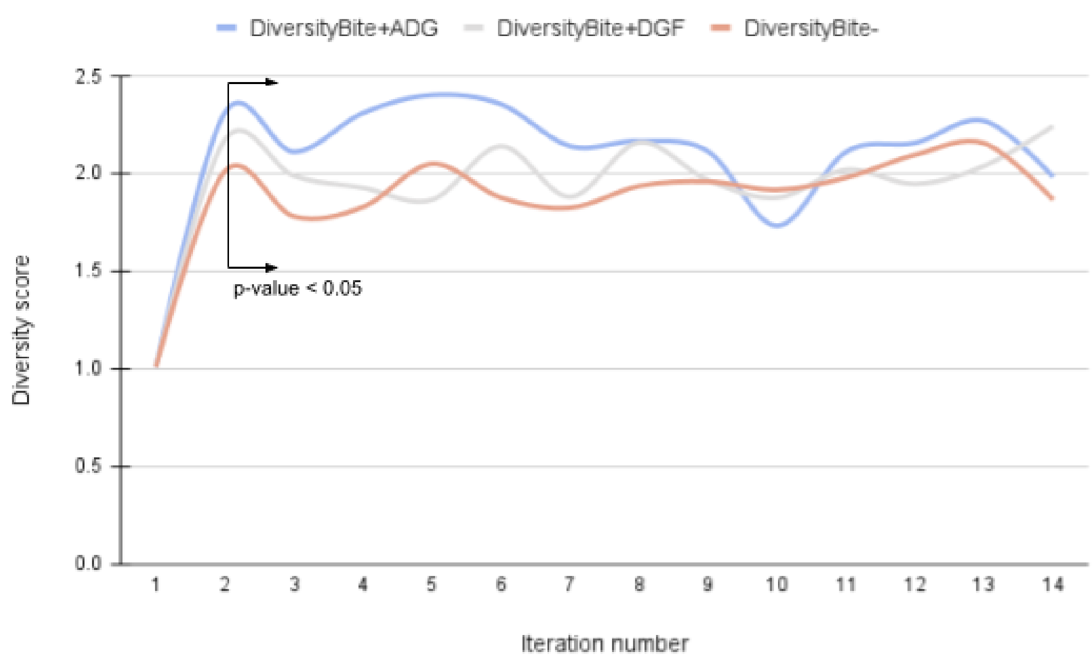


Figure 4.11: A comparison between DiversityBite-, DiversityBite+DGF, and DiversityBite+ADG diversity scores of the first 15 iterations for the same user

critique-based recommender has a higher diversity score. This finding suggests that the critique-based recommender can be applied in real scenarios even with a small number of interactions with the user.

4.5.4 Summary

This section presented a simulation study to compare between the two proposed generating critique dynamically algorithms and a baseline algorithm in recipe recommendation. The results show an improvement in the overall diversity score in comparison to a similarity-based baseline. This study provides a foundation for further research on diversity critique-based recommender approaches in the recipes domain. The next section provides the findings from conducting a controlled user study in incorporating diversity through the use of critique.

4.6 User Study: Evaluating Diversity in Dynamic Critique

4.6.1 Overview

The purpose of this user study is to answer the research question related to the impact of diversity-focused critique on user outcomes **RQ 2**. As a reminder **RQ 2** is restated here:

RQ 2: In critique-based conversational recommendation, how does diversity-focused critique impact diversity in terms of user outcomes?

- **H 2.1** Critique-based recommenders result in finding more diverse recipes compared to a non critique-based recommenders.
- **H 2.2** Users compile a more diverse meal plan in a critique-based recommender compared to non critique-based recommender.
- **H 2.3** Users perceive the modeled diversity of the recommended recipes.
- **H 2.4** In critique-based recommenders users prefer to explore using specific critique groups related to recipe features.

To address this research question, a full system of DiversityBite has been developed and a user study was conducted, in which users were asked to prepare a weekly meal plan by exploring recipes. The study includes participants of all ages and from different backgrounds and education levels. A user study was selected for this evaluation for its effectiveness in collecting real data logs of actual user behavior to aid in answering the research question.

4.6.2 Experiment Setup

A web-based recommender application of DiversityBite was developed for users to interact with. The user study was conducted to evaluate the effectiveness of using dynamic critique to recommend more diverse recipes. For the Top-N recommendation $\mathbf{N} = 10$, while the parameters to establish G-DGF was set to: $\mathbf{S} = 10$, $\mathbf{R} = 50$, $\mathbf{M} = 8$. The parameters were set through an empirical lab experiment to ensure reasonable computation time during user interaction with the system. To recall, $\mathbf{N}$ is the number of recommended recipes from the similarity-based retrieval model, $\mathbf{S}$ is the number of recipes used to establish G-DGF, $\mathbf{R}$ is the number of times the random selection process repeated, and $\mathbf{M}$ is the highest-percentage critique.

Two variations of DiversityBite were implemented: dynamic critique recommender (*Dynamic-Rec*), and static critique recommender (*Static-Rec*). A similarity-based recommender (*Sim-Rec*) was used as the baseline. *Dynamic-Rec* generates a dynamic critique list using G-DGF. *Static-Rec* displays the same set of critiques for each recipe, and finally, *Sim-Rec* is a similarity-based recommender. Recipes in this study were represented using the binary ingredients vector. In *Dynamic-Rec* and *Static-Rec*, the user explores more recipes by using the critique, while in *Sim-Rec* the user explores more recipes by navigating through several pages of recommended recipes. Given the close coupling between ingredients, flavor, and nutritional values, for this study, the ingredients were used to represent the recipes directly to calculate the diversity scores, while the flavor and nutritional features are used as critique options. Figure 4.12 shows the pre-survey and the reflection survey questions presented to the users as a part of their interaction with the application.

4.6.3 Demographic

The study included twenty-six participants that were recruited from students, staff, and faculty at a U.S. public university. The total duration time spent by each par-

Pre-survey questions		
#	Question	Options
1	For a dish you know, how often do you look for recipes?	Rarely/Sometimes/Often
2	When you prepare for a dish you know, how do you look for recipes?	Online/Asking relatives/Others
3	How often do you look for new recipes?	Rarely/Sometimes/Often
4	How do you look for new recipes?	Online/Asking relatives/Others
5	list some websites do you use?	Free text
6	What are the most important criteria do you look for when deciding on a recipes?	Free text
Reflection survey questions		
#	Question	Options
1	Did you find the recipe you were looking for?	Yes/No
2	Did you find new recipes?	Yes/No
3	Among the recipes you liked, are you willing to try one of them?	Yes/Some of them/None
4	Do you think recipes were similar to each other in each displayed list?	Yes, recipes were similar with small variations in ingredients
		No, recipes were different from each other
5	What was your main decision when you selected to see similar recipes?	Flavor
		Nutritional facts
		Preparation time & number of ingredients

Figure 4.12: Pre-survey and reflection survey questions along with available answers

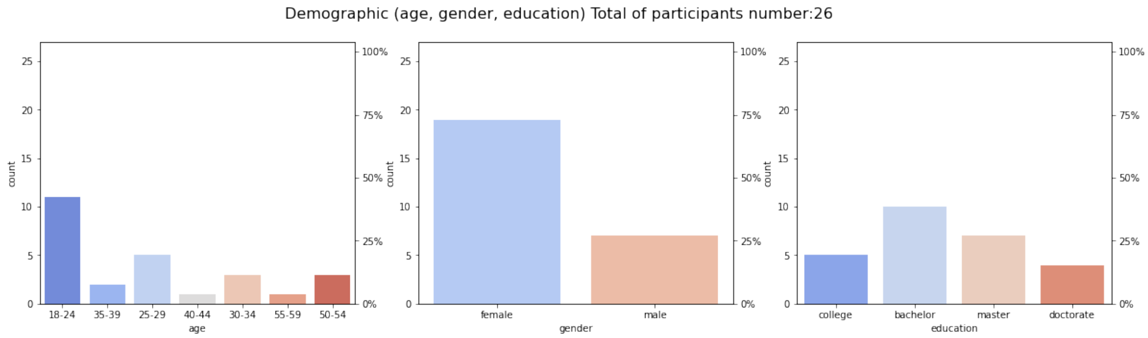


Figure 4.13: Demographic about participants age, gender, and education

ticipant was on average 40 minutes. Gender distribution was 19 females and 7 males. Most participants' age range was between 18 and 29 years and the majority of participants had at least a bachelor's degree. Figure 4.13 shows the distribution of age, gender, and education. All participants use 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. Figure 4.14, shows participants' responses to the first four questions in the pre-survey. Regarding the online resources they use, the most frequent resources are Google search, YouTube videos, and social networks. Recipe ingredients, preparation time, and balanced dish were the main criteria participants look for when deciding on a recipe. Figure 4.15, shows a wordcloud for the participants' text responses regarding question 5 and 6 in the pre-survey. This suggests that participants had a good exposure to online resources when looking for recipes. Among the chosen regions, Asia, Mediterranean, Middle Eastern, and North America were the most frequently chosen regions while the least chosen regions were Caribbean and Caucasus. The most frequently chosen meal courses were main dish, appetizers, and lunch while the least frequently chosen ones are beverages such as tea and cocktail. Figure 4.16, shows the distribution of participants selection of region, and course meal. On average participants spent around 8 minutes using each variation, and viewed on average 8 different recipe lists in each variation; 80 different recipes in each variation.

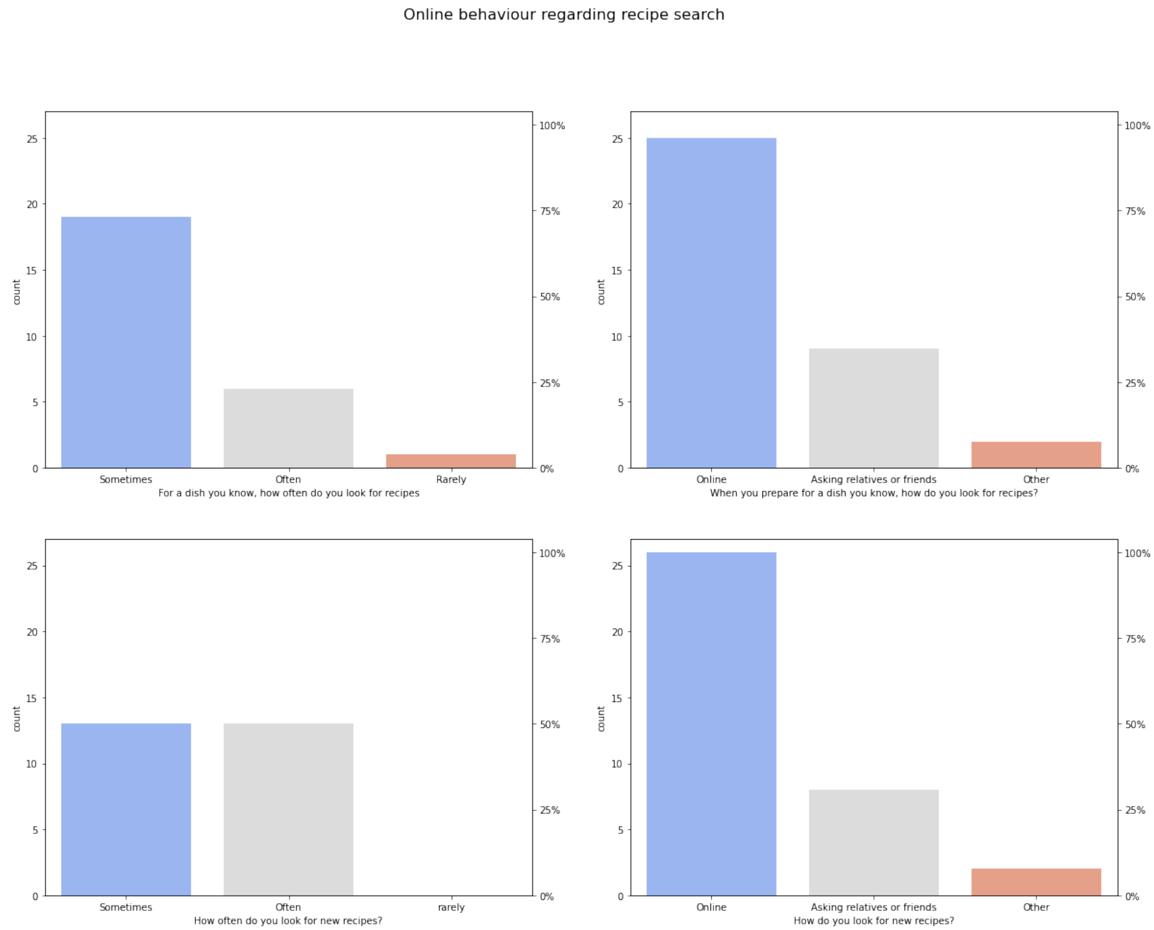


Figure 4.14: Distribution of Participants' responses for each question related to online behavior on searching for a new recipes

Figure 4.15: Wordclouds for answers provided by participants for Question 5 and Question 6.

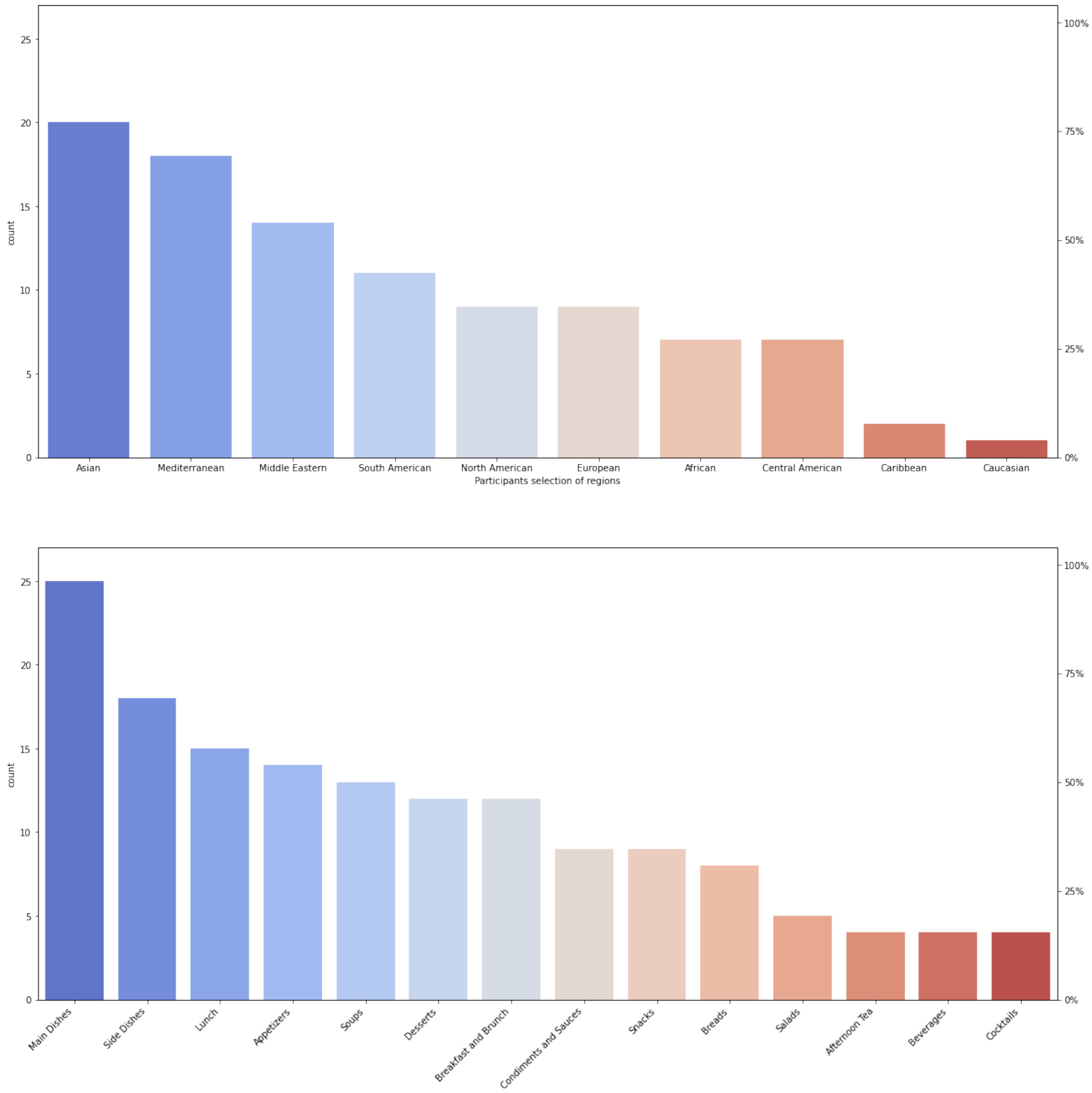


Figure 4.16: Participants selection of region and course meal

4.6.4 Analysis and Results

This section reports on finding related to the second research question by addressing the related hypotheses (H 2.1, H 2.2, H 2.3, H 2.4).

4.6.4.1 Meal Plan Size and Number of Disliked Recipes

To understand the effectiveness of the presented approach this dissertation looked at the meal plan size and number of disliked recipes. The intuition is that, given the same number of recommended recipes, the recommender system is more effective if the user can compile a larger meal plan while disliking fewer recipes. The rationale is that the more recipes the user disliked the more likely the recommender system was not able to satisfy the user's need. Figure 4.17, shows the average percentage of meal plan size to the number of recommended recipes (left), and the average percentage of disliked recipes to the total number of recommended recipes (right). The results in Figure 4.17 indicate that, participants added 20% of recommended recipes to their meal plan in *Sim-Rec* while only around 13% being added in the case of *Dynamic-Rec* and *Static-Rec*. For the number of disliked recipes, participants disliked 3% of recipes recommended by *Sim-Rec* and 2% in the case of *Dynamic-Rec* and *Static-Rec*. Users completed roughly the same number of iterations ($\sim$7.5) and spent the same amount of time ($\sim$7 minutes) in each recommender variation. A one-way repeated measure ANOVA test shows no statistical significance for the number of disliked recipes ($F(2,50) = 0.68$, p-value $= 0.51$), suggesting that all variations seemed to be equally similar in meeting participants' preference. However, for the meal plan size, there's a significant difference among the different variations ($F(2,50) = 3.79$, p-value < 0.05). Tukey's post hoc test shows that *Sim-Rec* had a significantly higher meal plan size than other variations. One attribute of this is the fact that the *Sim-Rec* recommender is driven by user preference and is more likely to generate recommendations that are compelling to the user and added to the meal plan. The similarity in the number of

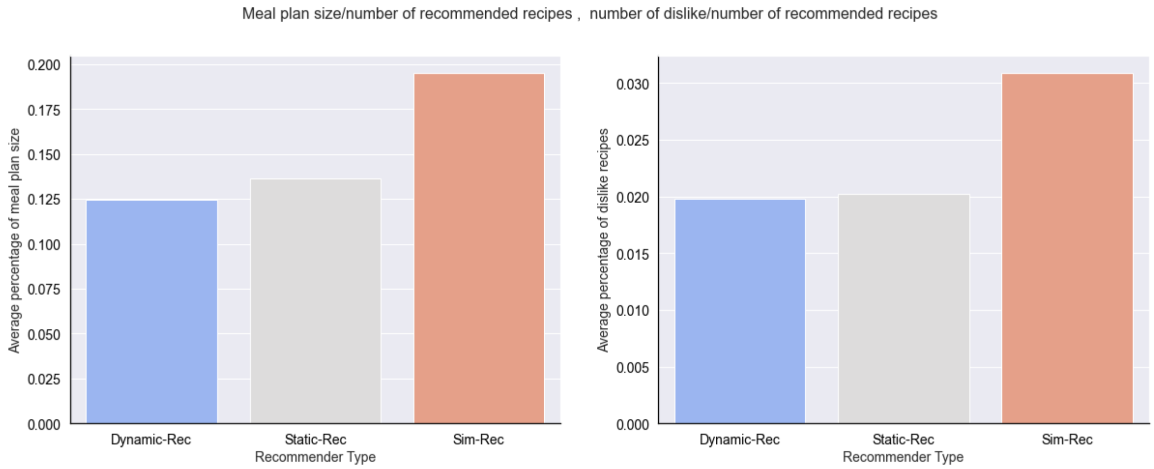


Figure 4.17: The average percentage of meal plan size to the number of recommended recipes (left), and the average percentage of disliked recipes to the total number of recommended recipes (right)

disliked items suggests that the ability to generate acceptable recommendations was not affected by the recommender variation.

4.6.4.2 Diversity in Meal Plan vs Diversity in recommended recipes

To address H 2.1 the diversity of recommended recipes at each iteration were analyzed. Figure 4.18 summarizes the average diversity scores for the recommended recipe per iteration for each variation. These results indicate that *Dynamic-Rec* has a higher diversity score compared to *Static-Rec* and *Sim-Rec*. A one-way repeated-measure ANOVA test shows a significant difference in the recommended recipe diversity ($F(2,50) = 42$, $p < 0.05$). Tukey's post hoc test shows that diversity in *Dynamic-Rec* is significantly higher than *Static-Rec* and *Sim-Rec*. It also shows that *Static-Rec* is significantly higher than *Sim-Rec*. Suggesting that, critiquing approaches were able to recommend more diverse recipes compared to non-critique approaches. Moreover, the diversity-focused critique was able to recommend more diverse recipes compared to a static critique. Therefore, H 2.1 is accepted.

To address H 2.2 the relationship between the diversity of the meal plan and the diversity of the recommended recipes were examined. The Meal plan diversity for each variation is summarized in Figure 4.19 (left). These results indicate that *Dynamic-Rec* has a higher diversity score compared to *Static-Rec* and *Sim-Rec*. A one-way repeated measure ANOVA test shows a significant difference in the meal plan diversity ($F(2,50) = 3.8$, $p < 0.05$). Tukey's post hoc test shows that diversity in *Dynamic-Rec* is significantly higher than *Sim-Rec*. Suggesting that, the dynamic critique approach was able to allow participants to create a more diverse list of meal plans and therefore H 2.2 is accepted. While the previous finding shows that participants created a larger meal plan in *Sim-Rec* compared to *Dynamic-Rec*, this finding shows that participants were able to create a more diverse meal plan in *Dynamic-Rec*.

Figure 4.19 (right) shows a scatter plot between the diversity of the meal plan on the horizontal axis and the average diversity score of recommended recipes. In all three variations, there is a direct relationship between the recommended recipes and the meal plan. There is a strong correlation between meal plan diversity and the diversity of the recommended recipes. The Pearson correlation for each variation is: *Sim-Rec* $r = 0.7$, $p < 0.05$, *Static-Rec* $r = 0.6$, $p < 0.05$, *Dynamic-Rec* $r = 0.4$, $p = 0.06$. The results suggest that for *Sim-Rec* and *Static-Rec* the diversity of recommended recipes affect the diversity of the meal plan diversity. While in the *Dynamic-Rec* the exploration process led the participant to a more diverse meal plan rather than the diversity in the recommended recipes.

4.6.4.3 Users behavior in Critique Selection

To validate H 2.4, participants' selection for critiques were analyzed in *Dynamic-Rec* and *Static-Rec*. Figure 4.20 shows that participants in *Dynamic-Rec* and *Static-Rec* have not shown preference for flavor critique over nutrition critique or vice versa. The figure also suggests that some critiques were less used compared to others such as Cholesterol, Trans Fat, Saturated Fat, and Bitterness. The reason behind that can be either (1) participants do not think about these features as important features to refine the search, or (2) participants were able to satisfy their needs by using other

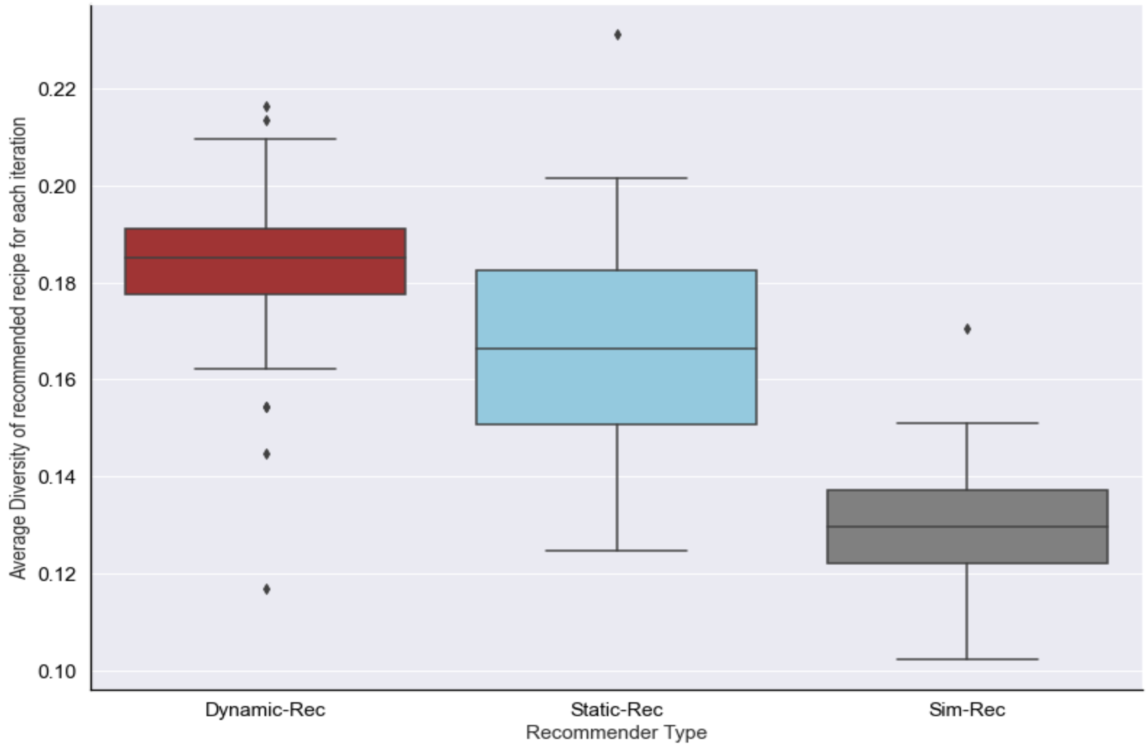


Figure 4.18: Average diversity score of recommended recipes per iteration for each recommender type

features that provide the same meaning. For example, in nutritional features, the Fat feature covers Trans Fat, and Saturated Fat as a critique. Another example can be sweetness is an inverse to bitterness and participants had more tendency to use sweetness instead of bitterness feature. It is worth noting that the protein critique was never been chosen in *Dynamic-Rec* variation while it was commonly chosen in *Static-Rec*, this can be explained that the protein feature will provide smaller steps towards more diverse recipes in the search space and therefore never been displayed to the participants as a critique option. As a result, participants used Meaty and Fat options more in *Dynamic-Rec* to compensate for the absence of protein critique.

To better understand the use of critique, figure 4.21 and figure 4.22 provide a breakdown for less/more critique distribution selected by participants in *Dynamic-Rec* and *Static-Rec*. Similar to the previous observation, there is no trend detected on preferring one group of features over the other. However, there is a similarity between

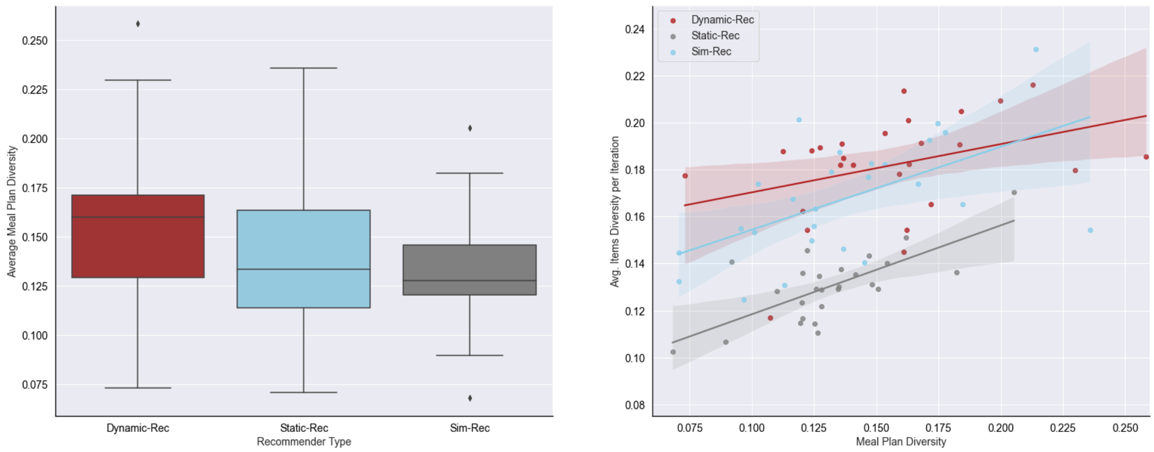


Figure 4.19: Meal plan diversity score for each variation (left), the relation between meal plan diversity and diversity of recommended recipes (right)

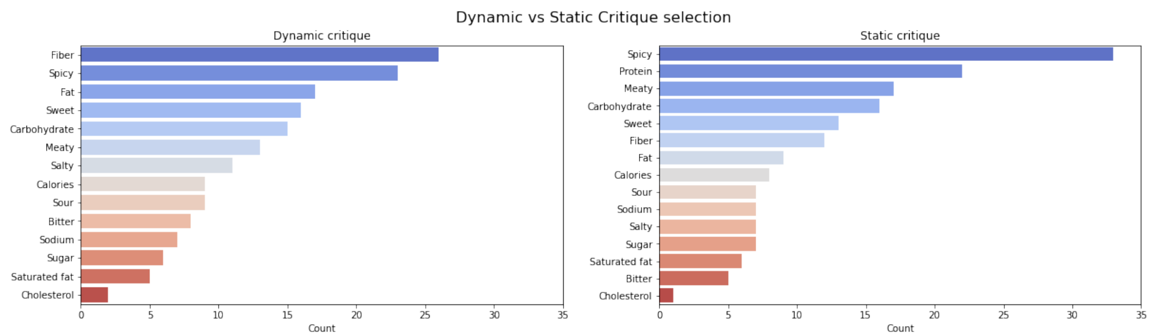


Figure 4.20: Most used critique (More/Less) in Dynamic and Static critiquing. The figure shows no user preference on using flavor critique or nutrition critique over the other.

Dynamic-Rec and *Static-Rec* critique selection in terms of direction (More/Less). For example, participants tend to use the 'less spicy' option more often in both *Dynamic-Rec* and *Static-Rec*.

These results do not provide support for hypothesis H 2.4 indicating that there was no user preference in using a group of features while exploring the recipes using critique-based recommenders.

4.6.4.4 Reflection Survey

The analysis of the reflection survey addresses H 2.3. Each participant had to answer a set of questions after using each recommender variation as shown in Table

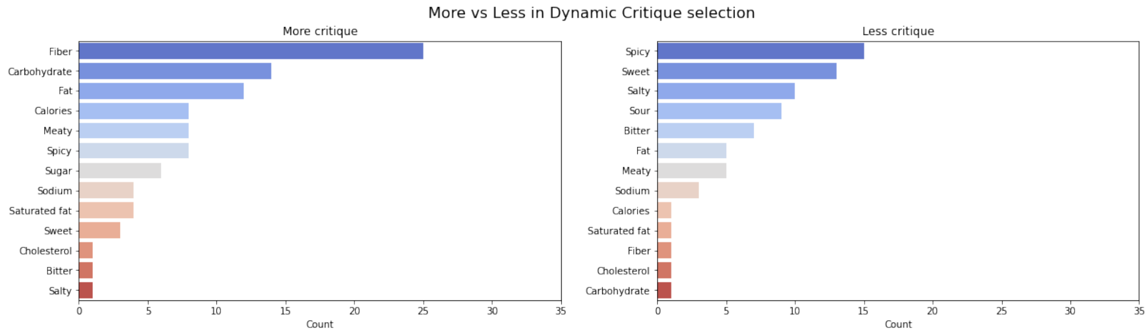


Figure 4.21: The frequency of selecting more critique vs less critique in *Dynamic-Rec*

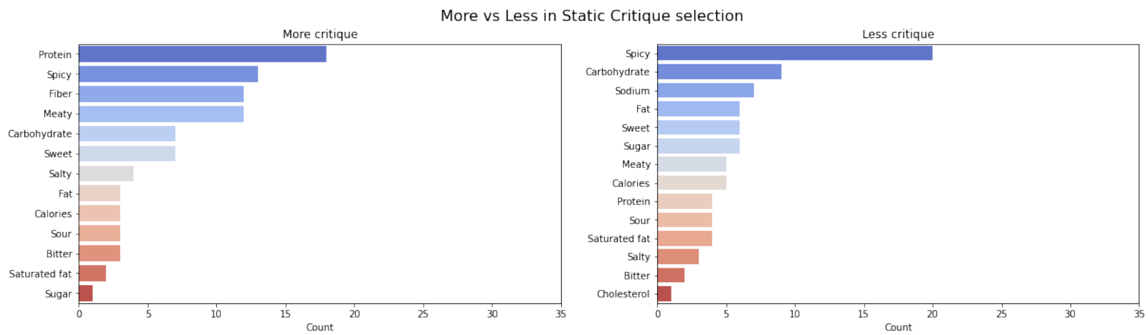


Figure 4.22: The frequency of selecting more critique vs less critique in *Static-Rec*

4.12 in the reflection survey part. For Q1 and Q2, all participants indicated that they found new recipes or found recipes they were looking for. In all variations, Q1 received on average fewer positive responses compared to Q2, Q1 received 19 positive responses compared to 25 for Q2. For Q3, none of the participants indicated in any variation that they will not try any of the recommended recipes. A statistical test for Q1, Q2, and Q3 shows no significant difference between all variations in recommending useful, novel, and valuable recipes. While the aim of this work is not to focus on novelty, usefulness, and valuable finding, the results of Q1, Q2, and Q3 provide an indication about the quality of the recommendation.

To capture participants' perception of diversity, in Q4 participants were asked if they thought recipes were similar to each other. According to [74], there are two types of diversity *categorical diversity* and *item-to-item diversity*. Q4 addresses the *item-to-item diversity* which aligns with the diversity style we are introducing. A chi-

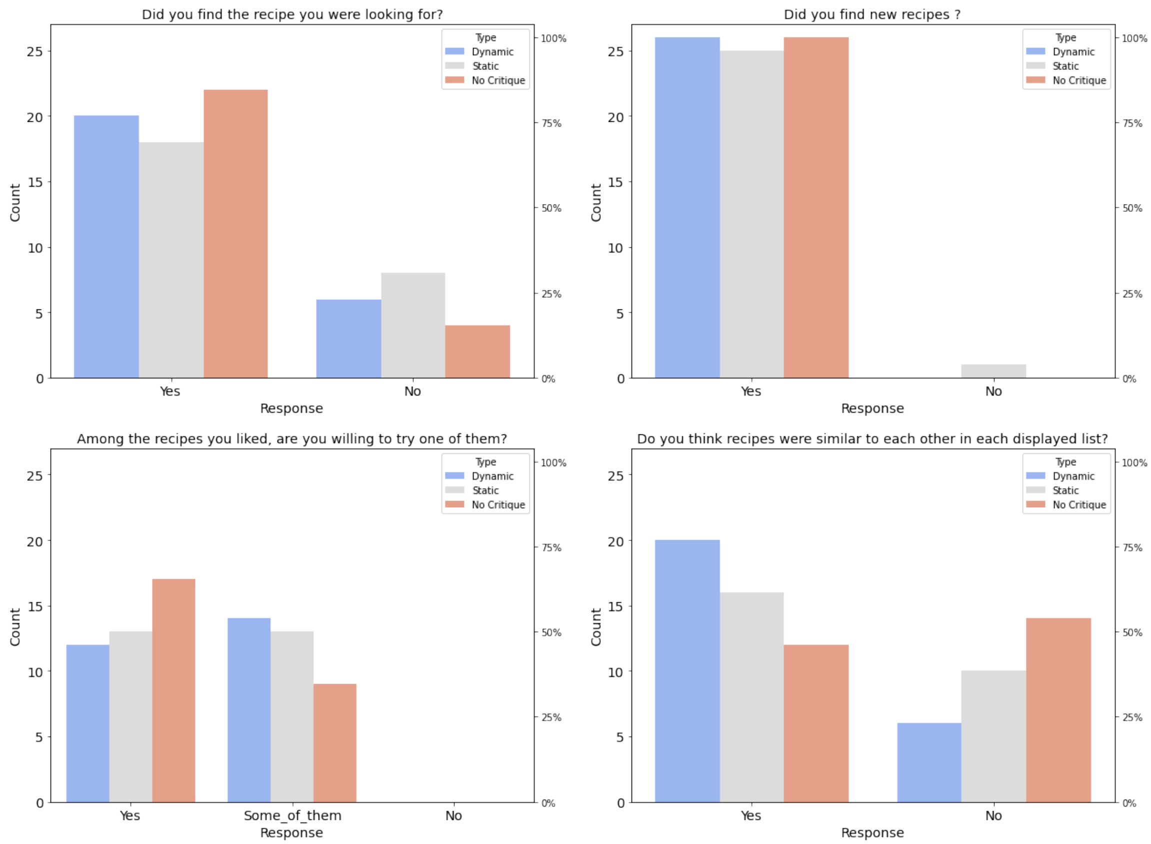


Figure 4.23: Participants answers to the reflection survey questions (Q1, Q2, Q3, and Q4

.

square test of independence showed that there was no significant association between recommender type and user perception of diversity $\chi^2(2, N = 26) = 5.2, p = 0.07$. This result aligns with the finding of [74] in which participants were not able to perceive item-to-item diversity. Therefore, introducing diversity in the recommendation while exploring will not make a noticeable difference to participants but can result in a diverse selection. Hence, H 2.3 was rejected. Figure 4.23 shows a distribution of participants' answers for Q1, Q2, Q3, and Q4.

The last question (Q5) in the post-survey asks participants about the main criteria they have used when selecting to see similar recipes. As shown in Figure 4.24 the results show mixed answers of Flavor, Nutritional facts, preparation time, and the number of ingredients. This aligns with participants' responses to the post-survey

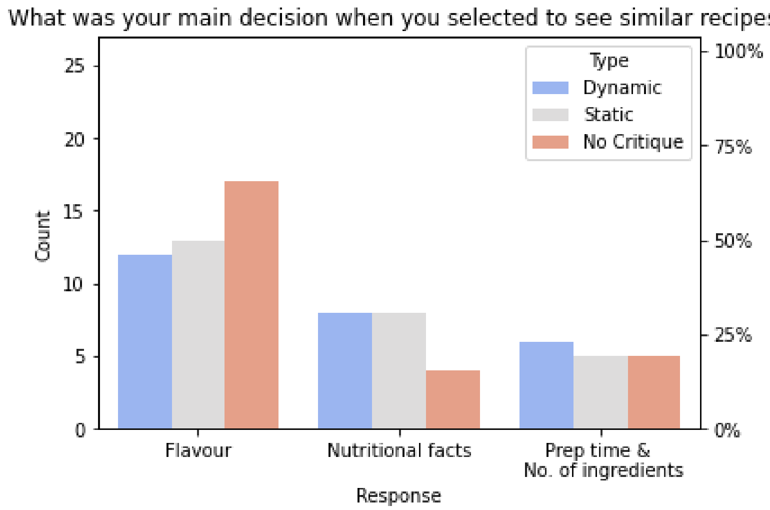


Figure 4.24: Participants answers to the criteria they followed to see similar recipes

.

question Q6. A chi-square test of independence showed that there was no significant association between recommender type and the criteria applied to see similar recipes $\chi^2(4, N = 26) = 2.75, p = 0.60$. This confirms the results of the user critique selection in *Dynamic-Rec* and *Static-Rec* conducted through log analysis. The logs show no clear distinction in the frequency of using flavor and nutrition critique. This finding suggests that participants were trying to utilize the available exploration option with no bias toward one type of critique.

4.6.5 Summary

The focus of this user study is to study the effectiveness of using diversity-focused critique on user outcomes in relation to diversity. Two variations of DiversityBite were implemented, *Dynamic-Rec*, and *Static-Rec*. In addition to a similarity-based recommender as a baseline. The user study answered the second research question (RQ 2), the results validated two hypothesis **H 2.1**, **H 2.2** and rejected the other two hypothesis **H 2.3**, and **H 2.4**. The main results show that participants were able to find more diverse recipes using critiquing methods compared to similarity-based retrieval **H 2.1**. Moreover, participants were able to compile a more diverse meal plan in dynamic critique compared to the other variations **H 2.2**. However, in all variations participants were not able to perceive the diversity of the recommended recipes **H 2.3**. Regarding **H 2.4**, there is no specific pattern that participants followed in selecting critique to explore.

4.7 User Study: Comparing between Different Recipe Representation

4.7.1 Overview

The purpose of this user study is to answer the research question related to the impact of recipe representation on diversity-focused critique recommendation in terms of user outcomes **RQ 3**.

RQ 3: In critique-based conversational recommendation, how does the underlying representation of recipes impact diversity in terms of user outcomes?

- **H 3.1** In diversity-focused critique, different recipe representation results in differences in the diversity of meal plans created by users.
- **H 3.2** In diversity-focused critique, recipe representation leads users to choose different types of critique features.
- **H 3.3** In diversity-focused critique, the diversity of the meal plan is realized based on different features that are related to certain demographic characteristics.
- **H 3.4** In diversity-focused critique, recipe representation impacts the user's perception of diversity.

To address this research question, a full version of DiversityBite has been developed and a user study was conducted, in which users were asked to prepare a weekly meal plan by exploring recipes. The study includes participants of all ages from different backgrounds and education levels. A user study was selected for this study because it is effective in collecting real data logs for actual user behavior to aid in answering the research question. This user study is similar to the previous user study in terms of user experience and task in compiling meal plan through the use of dynamic critique. However, they differ in terms of the critique generation algorithm and the recipe representation. The next section (Experiment Setup) discusses the details of the critique generation algorithm and the representation used in this user study.

4.7.2 Experiment Setup

A web-based recommender application of DiversityBite was developed for users to interact with. For details of the web application check Section 4.4. The user study was conducted to compare different types of underlying recipe representation while using dynamic critique. Adaptive Diversity Goal (ADG) was used to find the list of maximum diversity. For the Top-N recommendation $\mathbf{N} = 10$, while the parameters to establish ADG were set to: $\mathbf{S} = 10$, and $\mathbf{M} = 8$. The parameters were set through an empirical lab experiment to ensure reasonable computation time during user interaction with the website. To recall, **N** is the number of recommended recipes from the similarity-based retrieval model, **S** is the number of recipes used to establish ADG, and **M** is the number of critiques allowed per recipe.

Four variations of DiversityBite was implemented with four different representation, ingredient, flavor, nutrition, and nutrition flavor features: Ingredient representation (*Ingr-DiversityBite*), Flavor representation (*F-DiversityBite*), Nutrition representation (N-DiversityBite), and finally Nutrition and Flavor representation *FN-DiversityBite*. *Ingr-DiversityBite* was the baseline while the other three variations are the treatment. In all variations, the user explores more recipes by using the critique. Flavor and nutritional features were used as a critique for each variation regardless of the representation. Table 4.25 shows the pre-survey and the reflection survey questions.

4.7.3 Demographic

One hundred participants were recruited from students, staff, and faculty at a U.S. public university. The total duration time spent by each participant was on average 30 minutes. Gender distribution was 67 females and 33 males. Most participants' age range was between 18 and 24 years and the majority of participants had at least a bachelor's degree. Figure 4.26 shows the distribution of age, gender, and education. All participants use 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. Figure 4.27, shows participants' responses to the first four questions in the pre-survey. Regarding the online resources they use, the most frequent resources are Google search, YouTube videos, and social networks. Recipe ingredients, preparation time, and balanced dish were the main criteria participants look for when deciding on a recipe. Figure 4.28, shows a wordcloud for the participants who entered text regarding questions 5 and 6 in the pre-survey. 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 aligns with the demographic that participants were recruited from a U.S. public university. As the task was to plan 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. Figure 4.29, shows the distribution of participants' selection of region, and course meal. On average participants spent around 5 minutes using each variation, and viewed on average 7 different recipe lists in each variation meaning they were presented with at least 70 different recipes in each variation.

4.7.4 Analysis and Results

This section reports on the findings related to the third research question. The first section provides a general overview of each session, the subsequent sections address the research question.

4.7.4.1 Session Analysis

In all variations except the *Ingr-DiversityBite* variation, participants spent on average 5-6 minutes in each session. Further analysis has shown that the longer time spent in *Ingr-DiversityBite* is due to the computational time since recipes were represented using a higher vector dimension. Figure 4.30, shows the average time spent on each variation along with the computational time for each variation. In terms of exploration, participants were asked to explore recipes at least 7 times before being able to explore recipes in a different variation. Therefore, analysis shows that the majority of participants explored 7 times. However, as shown in Figure 4.31, some participants (shown on dots) explored more than 7 exploration iterations with Flavor having 20 explorations. For meal plan size, there was a similarity between variations. In addition to the exploration requirements, participants were asked to compile a meal plan of at least 7 recipes. Therefore, participants on average built meal plans of similar size. Figure 4.32, shows the average proportion of meal plan size to the number of recommended recipes in total. As shown in Figure 4.32, regardless of the variation, 14% on average of the recommended recipes were added to the meal plan. The website also allows participants to dislike recipes so that recipe will not be recommended in subsequent recommendation sessions. Participants disliked 5-7% of the recommended recipes. As shown in the figure, the most disliked recipes were recommended using the *NF-DiversityBite*.

4.7.4.2 Diversity in Meal Plan vs Diversity in recommended recipes

This section looks at five different types of diversity in recipes, diversity in terms of Ingredients, Flavor, Nutrition, Nutrition & Flavor, and finally Ingredients, Flavor, and Nutrition. Depending on the vector representation for the recipe. For example, when recipes were represented using the ingredients only the ingredient features were used in equation 2.1, therefore the calculated value of diversity represents the diversity in terms of ingredient.

Intuitively, the diversity of the meal plan depends on the diversity of the recommended recipes as participants created their meal plan from recommended recipes. Pearson correlation shows a direct relationship between both of them regardless of the diversity type or the representation type. Figure 4.33 shows the correlation between the meal plan diversity and the average diversity of the recommended recipes over the iteration. In the figure, each dot represents a participant, where the horizontal axis is the diversity of the compiled meal plan and the vertical axis is the average diversity of the recommended recipes over all iterations. Each participant had four dots representing each variation. Therefore, Figure 4.33 shows 400 dots.

Participants were able to create a more diverse meal plan in each variation depending on the definition of diversity. Table 4.2 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 score were highlighted. For example, using ingredient representation for recipes *Ingr-DiversityBite* participants created meal plans with the highest diversity in ingredients. Similarly, participants created a meal plan with high diversity in terms of nutrition using *Ingr-DiversityBite* variation. The table also shows that using *N-DiversityBite* variation participants created diverse meal plans 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. These results show that different representation yields differences in the meal plan diversity scores.

For ingredients, a one-way repeated measure ANOVA test shows there's a significant difference between the diversity scores in the meal plan for each variation ($F(3,297)=6.82$, $p < 0.05$). Tukey's post hoc test shows that the 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 there's 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 the 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 there's 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 the diversity score in *F-DiversityBite*, is significantly higher than the diversity scores in *FN-DiversityBite*. Finally, for all features, a one-way repeated measure ANOVA test shows a significant difference between the diversity scores in the meal plan for each variation ($F(3,297)=5.64$, $p < 0.05$). Tukey's post hoc test shows that the 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 the results show that diversity in meal plans differs depending on the representation which supports hypothesis H 3.1.

To summarize, the results show that using ingredient representation participants were able to create meal plans that are diverse in ingredients, and nutrition. While using nutrition and flavor representation participants created a diverse meal plan in terms of nutrition and flavor. Finally, using nutrition only representation participants compiled a diverse meal plan with respect to all features (ingredient, nutrition, and flavor). We note here again that the four variations were counterbalanced to reduce the order effect among the 100 participants. One-way ANOVA shows that there are

Table 4.2: Average meal plan diversity for each variation along with the definition of diversity

Diversity Type	Variation	Average Meal Plan Diversity
Ingredients	*Ingr-DiversityBite*	0.0951034
Ingredients	*F-DiversityBite*	0.08188447
Ingredients	*N-DiversityBite*	0.09154105
Ingredients	*FN-DiversityBite*	0.09296638
Nutrition	*Ingr-DiversityBite*	0.39415744
Nutrition	*F-DiversityBite*	0.38170294
Nutrition	*N-DiversityBite*	0.39173843
Nutrition	*FN-DiversityBite*	0.38722475
Flavor	*Ingr-DiversityBite*	0.39367217
Flavor	*F-DiversityBite*	0.39546474
Flavor	*N-DiversityBite*	0.39568812
Flavor	*FN-DiversityBite*	0.39435912
Nutrition & Flavor	*Ingr-DiversityBite*	0.37585146
Nutrition & Flavor	*F-DiversityBite*	0.3794536
Nutrition & Flavor	*N-DiversityBite*	0.37236971
Nutrition & Flavor	*FN-DiversityBite*	0.36656248
All features	*Ingr-DiversityBite*	0.12482654
All features	*F-DiversityBite*	0.13438651
All features	*N-DiversityBite*	0.13483653
All features	*FN-DiversityBite*	0.13413621

no statistical significant differences between the four variations in terms of the order when it comes to the diversity of the recommended recipes. Meaning that the diversity scores of the first variation is not significantly different from the second and so on for other variations. This also applies to the composed meal plan, in which one-way ANOVA shows that no statistical significant differences between the four variations in terms of the order regarding composing a meal plan.

4.7.4.3 Users behavior in Critique Selection

This section addresses hypothesis H3.2 which is related to participants' behavior on selecting critique. Critiques can be viewed into two main types: Nutrition critique, and Flavor critique. To recall, nutrition critiques are Protein, Calories, Carbohydrates, Sugar, Fiber, and Fat. On the other hand flavor critique are Bitter, Sour,

Table 4.3: The results of t-test statistical significance of critique types selection by participants

Variation	t-value	DoF	P-value
Ingr-DiversityBite	5.93	198	< 0.05*
N-DiversityBite	9.18	198	< 0.05*
F-DiversityBite	0.81	198	0.42
NF-DiversityBite	6.60	198	< 0.05*

Salty, Meaty, Spicy, and Sweet. Figure 4.34, shows that participants preferred to explore using flavor critique rather than nutrition critique in all variations except in flavor representation (*F-DiversityBite*). This suggests that *F-DiversityBite* was able to provide participants recipes with flavor that matches preference and therefore participants chose nutrition critique to explore more. Table 4.3, shows the results of t-test. These results support hypothesis H3.2 in which representation types can lead users to prefer one type of critique over the other. In *F-DiversityBite* variation recommended recipes that match participants' flavor interest but not nutrition. Therefore, participants chose to explore recipes using nutrition features. This result confirms participants' responses to Q7 in the reflection survey: "What influenced your main decision when you explored more recipes?". Figure 4.35, shows a summary for the participants response. The figure shows that in all variations participants chose flavor as the main reason that influenced the decision to explore. To check if there are any order effects in terms of users selection of critique, one-way ANOVA shows no significant differences between the orders in users selection of flavor and nutrition critique.

4.7.4.4 Demographic Differences in Diversity and Critique Selection

To address hypothesis H 3.3, participants were split into two groups depending on gender. In terms of meal plan diversity, male participants created a more diverse meal plan using nutrition representation, while female participants created a more diverse meal plan using flavor representation. One-way ANOVA with repeated measures shows a significant difference in the 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 *F-DiversityBite* variation are significantly higher than the diversity scores in *FN-DiversityBite*. For males, one-way ANOVA with repeated measures shows a statistical significance 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 than the diversity scores in *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 the meal plan in terms of flavor. To further explore this observation, critique selection for males and females was analyzed.

Critique selection analysis shows that females always chose to explore more recipes using flavor critique compared to nutrition. On the other hand, male participants used the flavor and nutrition critique in two representations equally. Figure 4.36 and 4.37, shows a comparison between flavor and nutrition critique between male and female groups. This finding supports hypothesis H3.3 in which demographic differences show differences in terms of meal plan diversity and critique behavior selection.

Regarding other demographic features, the analysis shows no statistical differences between groups in terms of age and education. Therefore, the analysis introduced in this section was focused on the differences between males and females.

4.7.4.5 Reflection Survey

The analysis of the reflection survey addresses hypothesis H3.4. Each participant had to answer a set of questions after using each recommender variation as shown in Table 4.25 in the reflection survey part. Figure 4.38, shows a screenshot of the reflection survey displayed for users during the user study. For Q1 and Q2, all participants indicated that they found new recipes or found recipes they were looking for. In all variations, Q1 and Q2 received on average a similar number of ratings on a 5-point

Likert scale with 1 as strongly disagree and 5 strongly agreeing. A statistical test for Q1, and Q2 shows no significant difference between all variations in recommending useful, novel, and valuable recipes. While the aim of this work is not to focus on novelty, usefulness, and valuable finding but the results of Q1, and Q2 provide an indication about the quality of the recommendation.

To capture participants' perception of diversity, participants in Q3, Q4, and Q5 were asked questions addressing the variety of the recommended recipes and the created meal plan. According to [74], there are two types of diversity *categorical diversity* and *item-to-item diversity.* In the recipes domain, categorical diversity refers to diversity in terms of cuisine and course meal while item-to-item diversity refers to differences between the recipes such as ingredient, nutrition, and flavor. In the three questions, the terms variety and different were used instead of diversity to avoid priming users into thinking about diversity. Q3 (I have seen recipes of different varieties), and Q5 (I was able to create a meal plan from different varieties of recipes) addresses categorical diversity. Q4 (Recipes in my meal plan were similar to each other) addresses item-to-item to diversity. Figure 4.39, shows the combined average results of participants responses for all questions. As shown in the figure the average rating in all questions was around 3.5. In all variations, none of the questions were statistically significant. This result aligns with previous findings in literature such as [74] and the previous experiment explained before. Therefore, introducing diversity in the recommendation while exploring will not make a noticeable difference to participants but can result in a diverse selection.

4.7.5 Summary

The focus of this user study is to explore the effect of the underlying representation of the recipe on user outcomes in relation to diversity while using diversity-focused critique. Four variations of DiversityBite were implemented, *Ingr-DiversityBite*, *F-DiversityBite*, *N-DiversityBite*, and *FN-DiversityBite*. Each variation uses different representations of recipes: Ingredients, Flavor, Nutrition, Flavor, and Nutrition respectively. All variations use the same similarity component and critique-generation component. The user study answered the third research question (RQ3), the results validated the first three hypothesis **H 3.1**, **H 3.2** and **H 3.3** and rejected the last hypothesis **H 3.4**. The main findings show that choosing a specific set of recipe representations will increase the diversity of meal plans in terms of a specific dimension. For example, using nutrition representation increases the ingredient, nutrition, and flavor diversity of the meal plan **H 3.1**. Moreover, participants tend to use flavor critique more than nutrition critique in all variations except in *F-DiversityBite* **H 3.2**. This study also shed light on the demographic differences in creating meal plans and using critique **H 3.3**. Finally, as shown in other studies participants were not able to perceive diversity in any variation. This finding suggests that the used variation will not allow the user to notice any diversity on the recommended recipe or the created meal plan therefore hypothesis **H 3.4** is rejected.

4.8 Discussion

This chapter presents three studies, the first study is a simulation study and the other two studies are user studies. In all studies, different variations of *DiversityBite* were evaluated to address the research questions. The results of each study have provided answers to the research questions presented in this dissertation.

First, a simulation study is conducted to understand the diversity scores of recommended recipes over the course of user interaction. In addition to understanding the feasibility of the critique generation algorithm in providing diverse recipes over the course of interaction. The results confirmed that diversity can be increased using the proposed critique generation technique. Moreover, the results show that the algorithm can be applied in the real domain by providing a diverse critique after the third iteration. The results of the simulation study serves as a basis to a more comprehensive experiment. The results show that exploration using critique recommends more diverse items when compared to non-critique approach.

Second, a user study is conducted to study the impact on diversity through the use of diversity-focused critique in terms of user outcomes. The result shows that critique-based recommenders recommend more diverse recipes compared to non-critique-based recommenders. It also enabled users to create more diverse meal plans as well. The study also shows that neither in critique-based recommender nor in non-critique-based recommender participants aren't able to perceive the diversity of the recommended recipes. Finally, the results illustrate that participants don't follow a specific pattern in selecting critique while exploring recipes to create a meal plan. The findings in this user study confirms with the finding in the simulation study in which the critique-based approach recommends more diverse recipes compared to non-critique-based approach.

Finally, the last user study is conducted to study the differences between using different recipe representations on diversity in critique-based recommenders. The analysis shows that recipe representation affects the diversity of meal plans created by participants. In addition to that, different recipe representations lead users to choose different critiques. These two findings were also extended to show differences between demographic groups. Finally, similar to previous studies, user perception of diversity was very minimal and not statistically significant between different recipe representations. This study suggests that diversity in terms of specific dimension can be increased by using a different recipe representation. This finding is helpful in scenarios where diversity is needed to be increased in one dimension over the other such as diversity in nutrition over diversity in flavor.

Pre-survey questions		
#	**Question**	**Options**
1	For a dish you know, how often do you look for recipes?	Rarely/Sometimes/Often
2	When you prepare for a dish you know, how do you look for recipes?	Online/Asking relatives/Others
3	How often do you look for new recipes?	Rarely/Sometimes/Often
4	How do you look for new recipes?	Online/Asking relatives/Others
5	list some websites do you use?	Free text
6	What are the most important criteria do you look for when deciding on a recipes?	Free text
Reflection survey questions		
#	**Question**	**Options**
1	I found new recipes	(1-5) Strongly agree – Strongly disagree
2	Among the recipes I liked, I am willing to try all of them	
3	I have seen recipes of different variety	
4	Recipes in my meal plan were similar to each other	
5	I was able to create a meal plan from different variety of recipes	
6	The waiting time for the new recipes to load was acceptable	
7	What influenced your main decision when you explored more recipes?	Flavor Nutritional facts Preparation time & number of ingredients

Figure 4.25: Pre-survey and reflection survey questions along with available answers

.

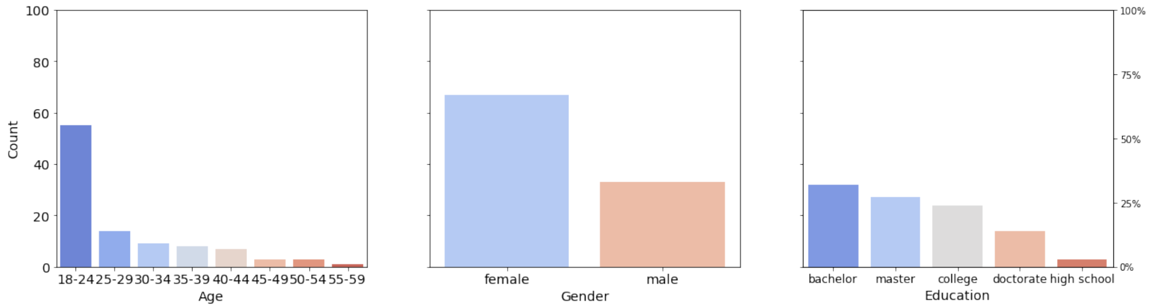


Figure 4.26: Demographic for the 100 participants

.

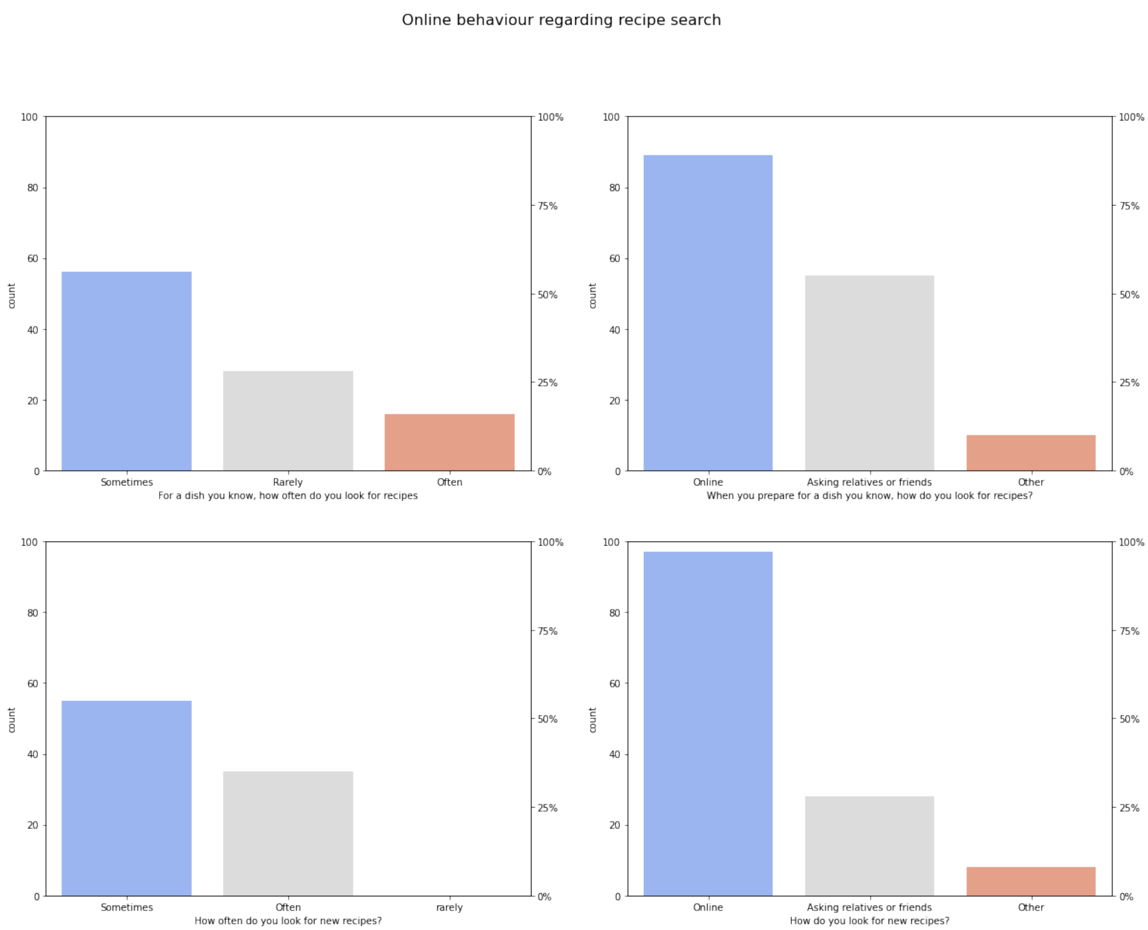


Figure 4.27: Demographic for the 100 participants

.

Figure 4.28: Wordclouds for answers provided by participants for Question 5 and Question 6.

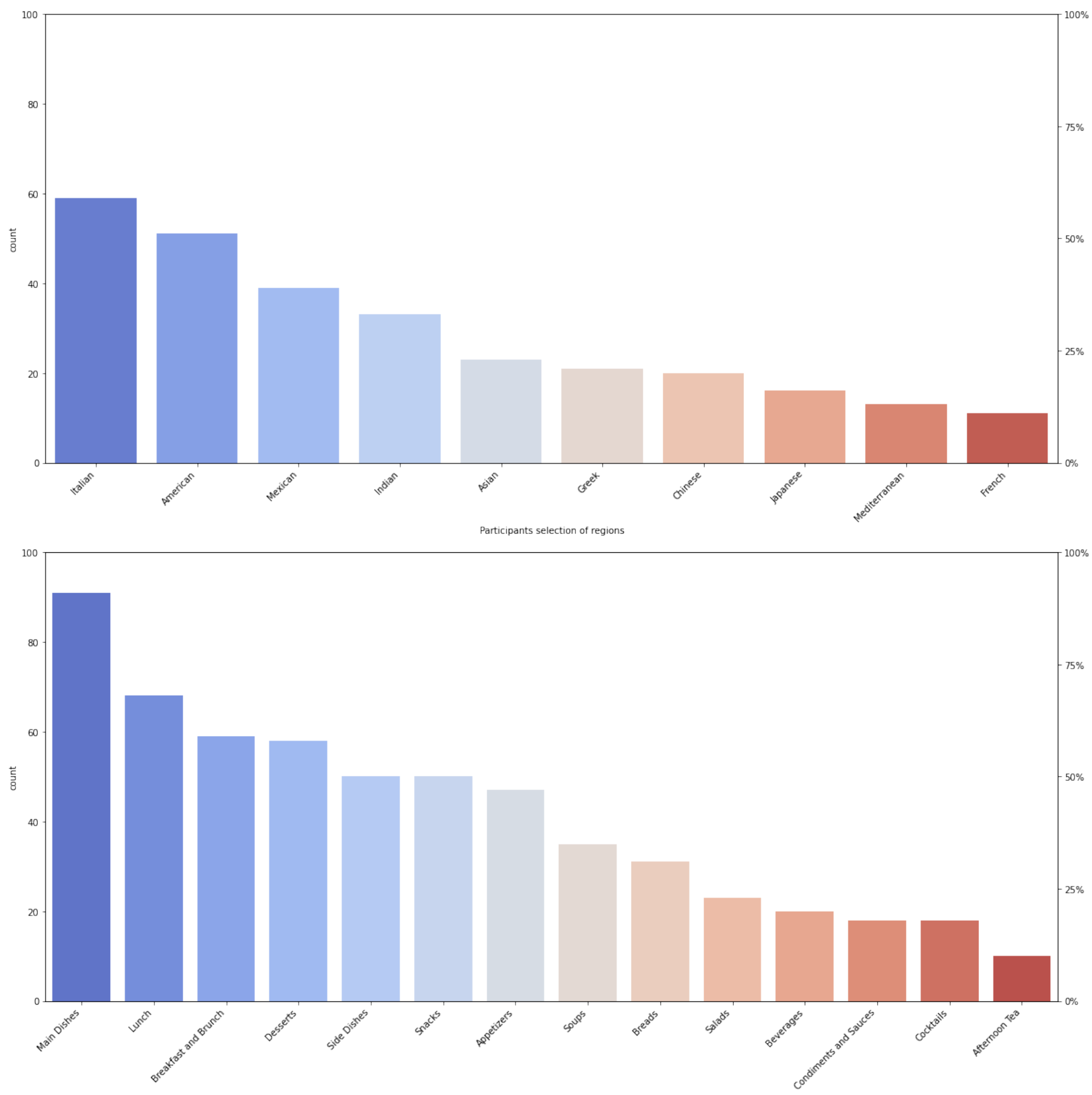


Figure 4.29: Participants selection of region and course meal

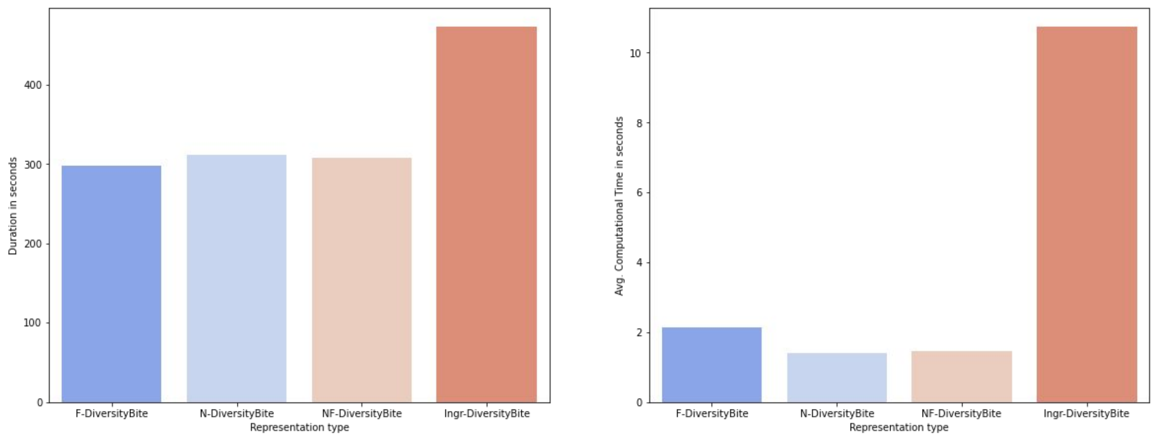


Figure 4.30: The average duration spent on each variation in seconds (left), the average computational time for each variation (right)

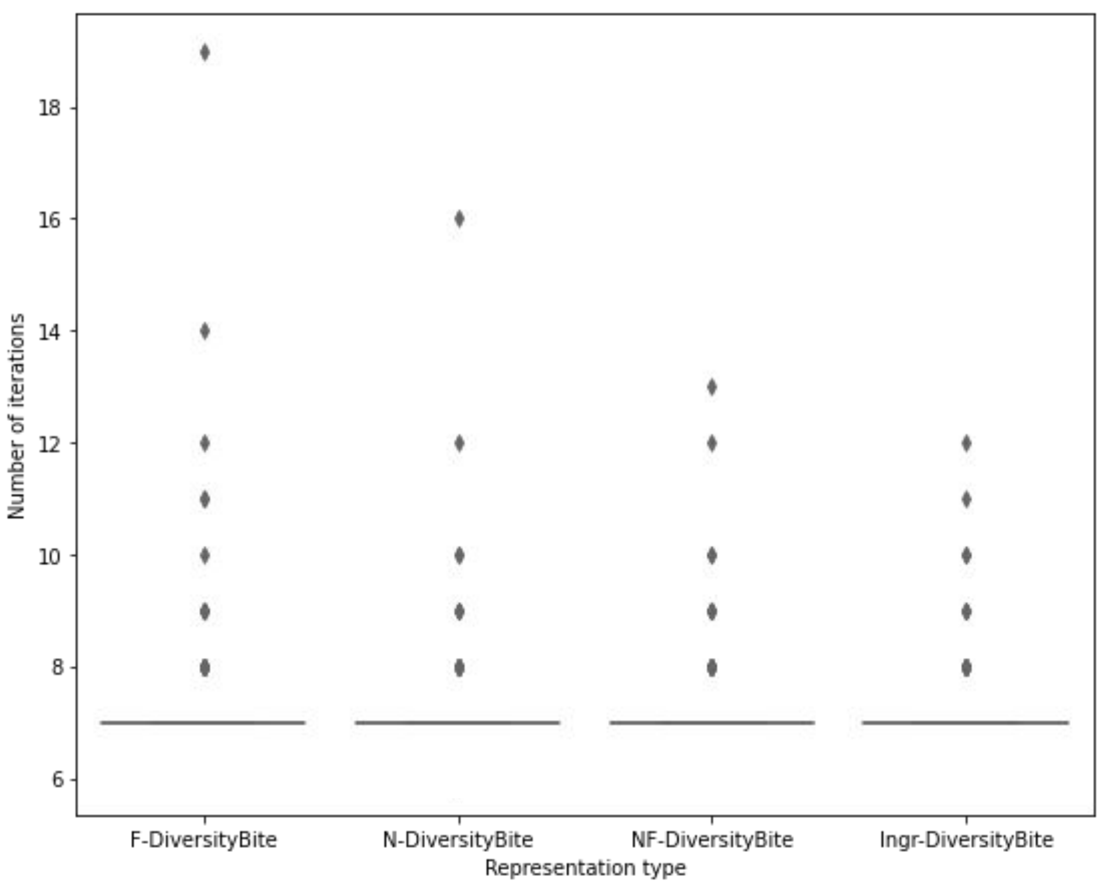


Figure 4.31: The number of iterations for each variation

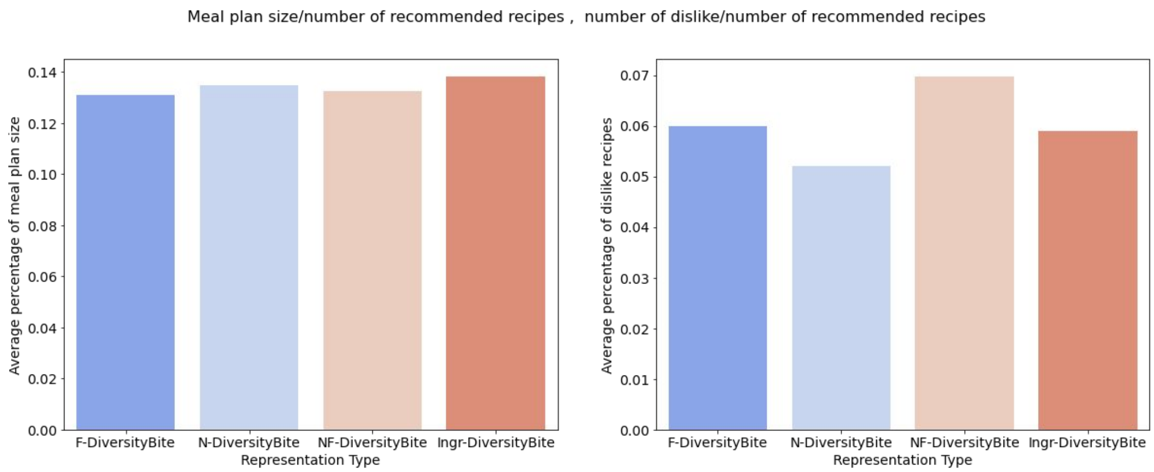


Figure 4.32: Meal plan size, and number of disliked recipes in proportion to the total number of recommended recipes

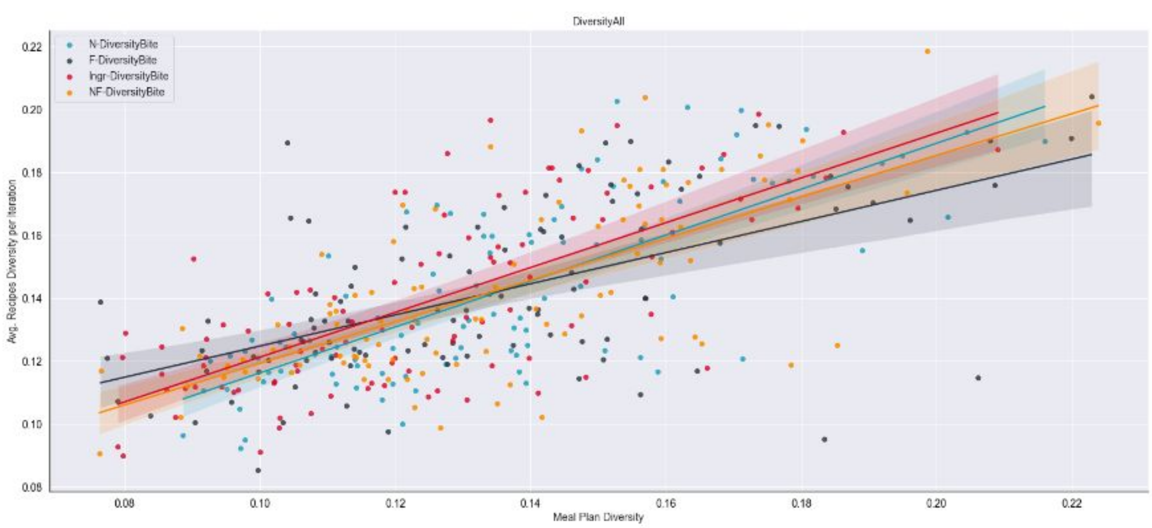


Figure 4.33: The number of iterations for each variation

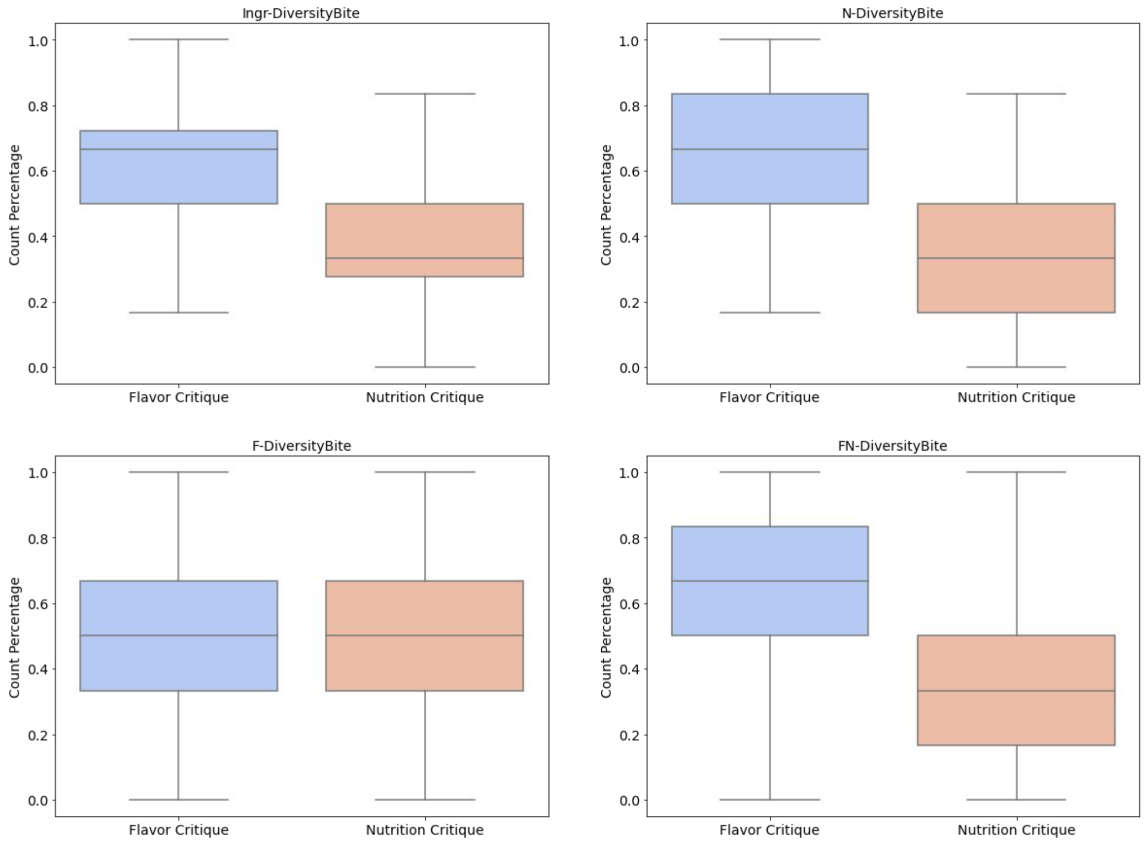


Figure 4.34: A comparison between participants selection on flavor and nutrition critique on different variations

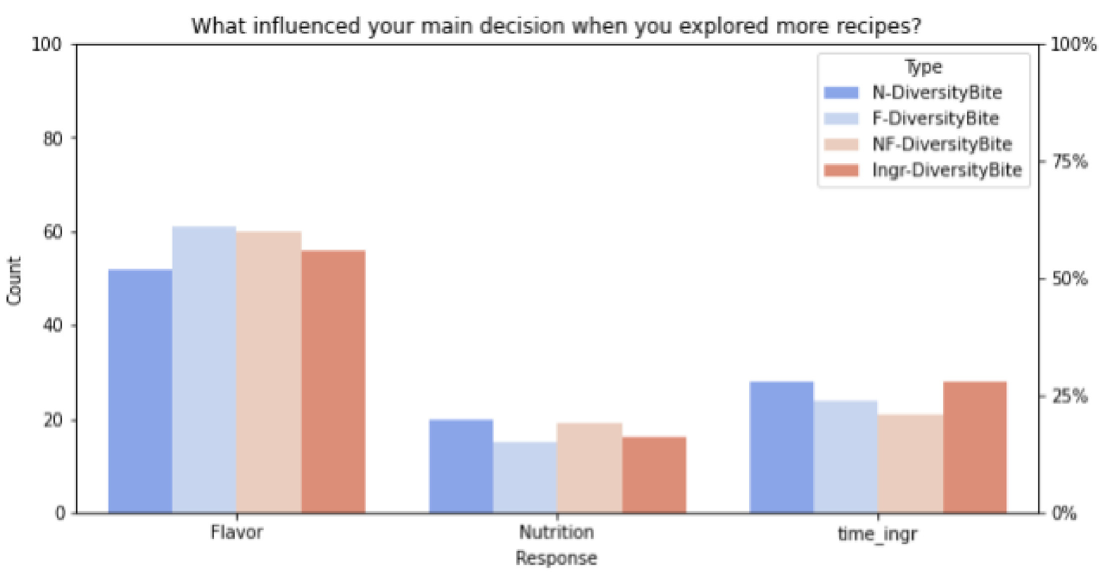


Figure 4.35: A comparison between participants responses for Q7 in the reflection survey

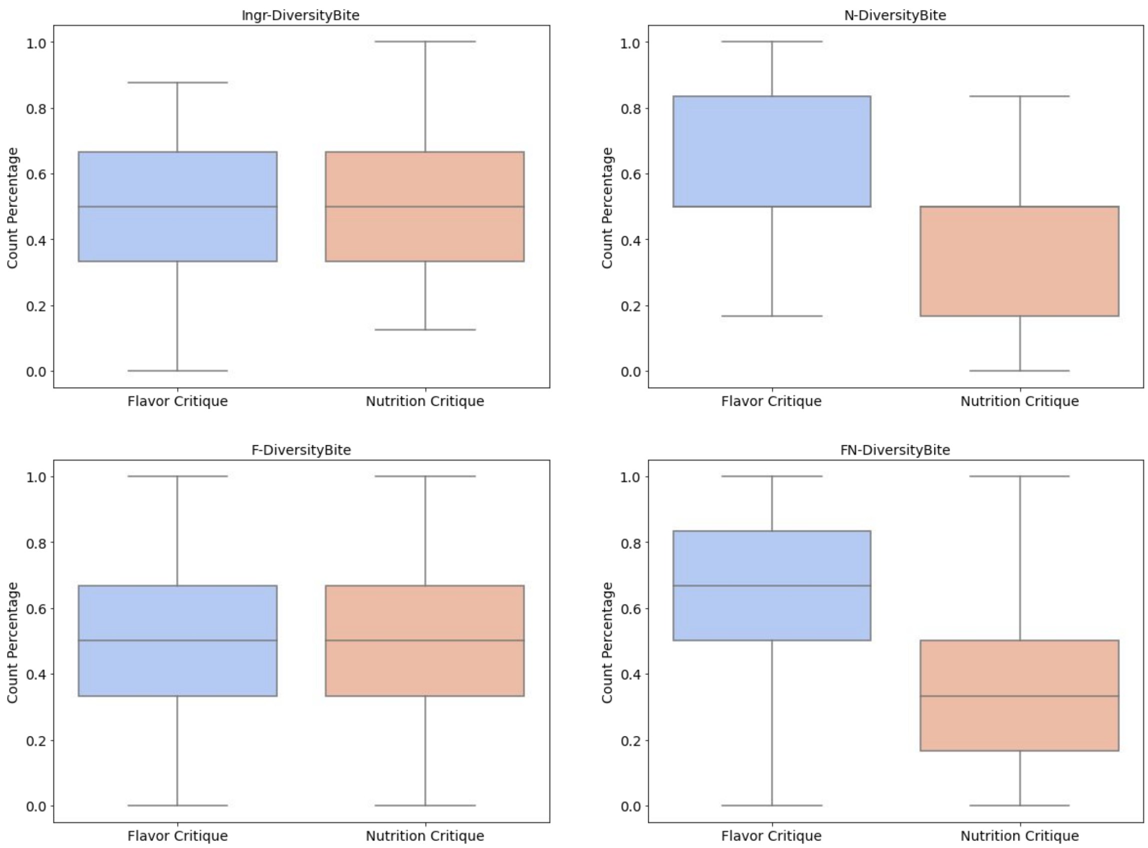


Figure 4.36: Critique selection percentage for male participants in each variation

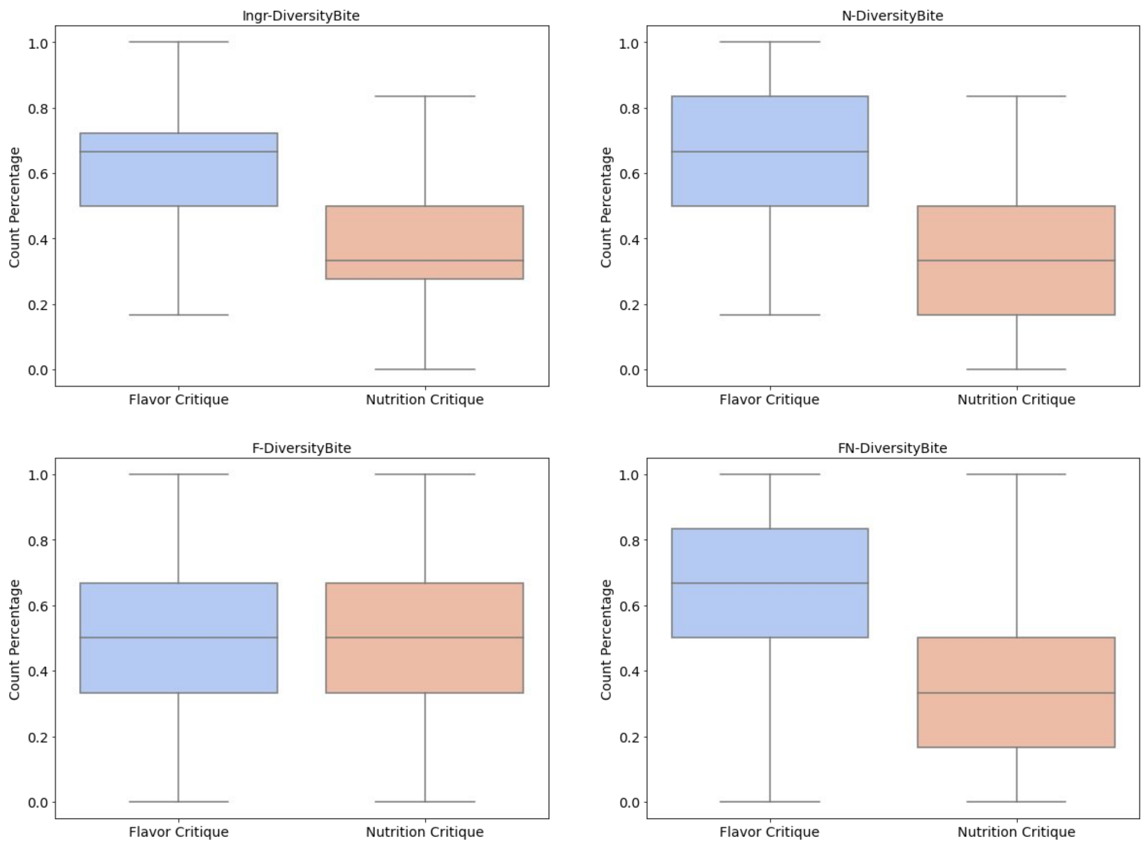
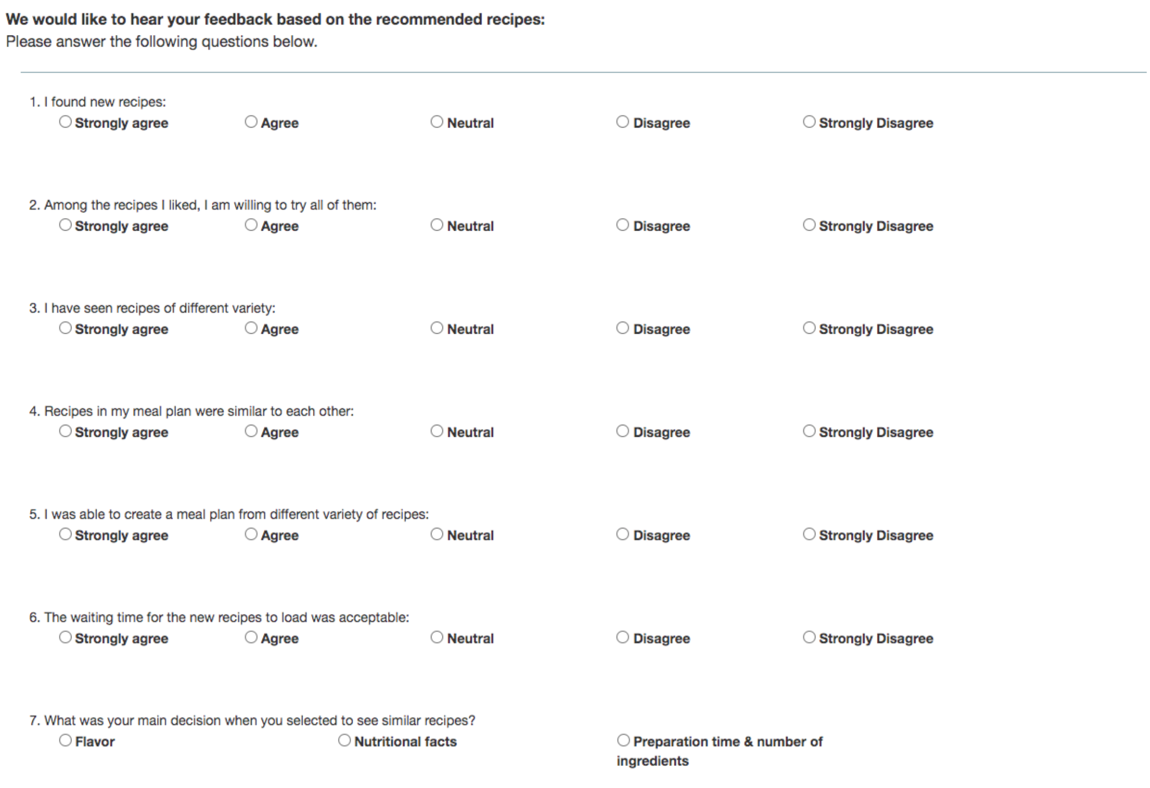


Figure 4.37: Critique selection percentage for female participants in each variation



We would like to hear your feedback based on the recommended recipes:
Please answer the following questions below.

1. I found new recipes:
○ **Strongly agree** ○ **Agree** ○ **Neutral** ○ **Disagree** ○ **Strongly Disagree**

2. Among the recipes I liked, I am willing to try all of them:
○ **Strongly agree** ○ **Agree** ○ **Neutral** ○ **Disagree** ○ **Strongly Disagree**

3. I have seen recipes of different variety:
○ **Strongly agree** ○ **Agree** ○ **Neutral** ○ **Disagree** ○ **Strongly Disagree**

4. Recipes in my meal plan were similar to each other:
○ **Strongly agree** ○ **Agree** ○ **Neutral** ○ **Disagree** ○ **Strongly Disagree**

5. I was able to create a meal plan from different variety of recipes:
○ **Strongly agree** ○ **Agree** ○ **Neutral** ○ **Disagree** ○ **Strongly Disagree**

6. The waiting time for the new recipes to load was acceptable:
○ **Strongly agree** ○ **Agree** ○ **Neutral** ○ **Disagree** ○ **Strongly Disagree**

7. What was your main decision when you selected to see similar recipes?
○ **Flavor** ○ **Nutritional facts** ○ **Preparation time & number of ingredients**

Figure 4.38: A screenshot of the reflection survey questions.

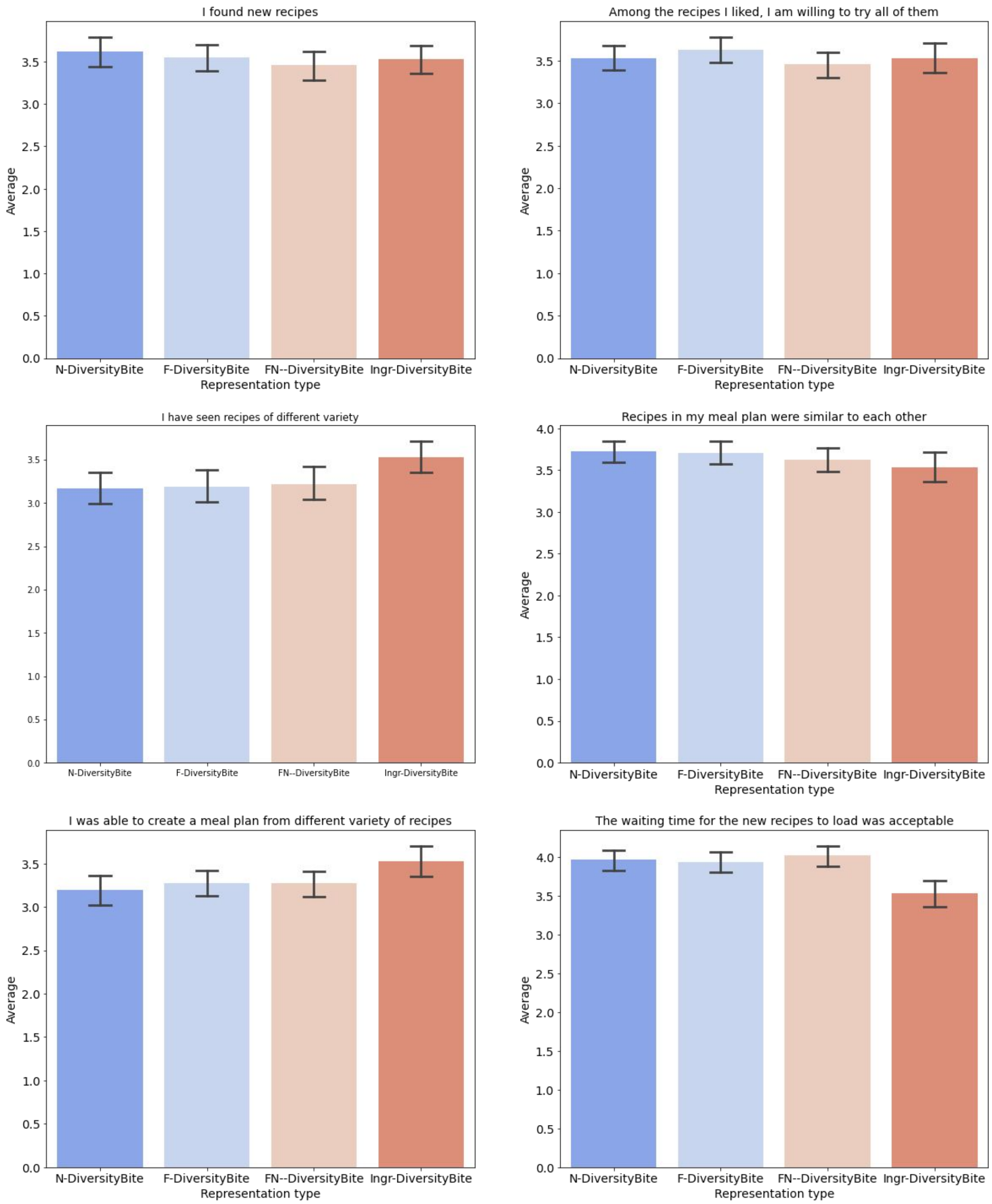


Figure 4.39: Reflection survey results for all participants. The figure shows the average score for each variation

CHAPTER 5: Future Work and Conclusion

5.1 Overview

This chapter summarizes the findings and contributions of this dissertation, as well as the limitations, direction for future research, and conclusion.

5.2 Contributions

The major contributions of this dissertation are three-folds. First, a novel approach of dynamic diversity-focused critique for the conversational recommender system was developed as introduced in Chapter 3. Second, the dynamic diversity-focused critique applied in the recipes domain to support diet diversification while exploring as shown in the evaluation study in section 4.6. Finally, the identification of recipe features that help find diverse recipes using dynamic critique as shown in the second evaluation study in section 4.7.

In Chapter **??**, we introduced a novel approach of dynamic critique. This dissertation implemented two variations of the proposed approach Diversity-Goal Footprint (DGF) and Adaptive Diversity-Goal (ADG). DGF and ADG are dynamic critique approaches in which critiques were generated from the space of all available critiques. The dynamic critique approach is diversity-driven in which the selected critiques were selected to increase the overall diversity of recommended recipes. In a way that can be thought of as a "like this, but more diverse". Unlike other critique-based CRS that tries to increase similarity through fine-tuning results using critique, this dissertation uses critique to guide users into more exploration. The results of the simulation studies show the feasibility of both computational approaches and that they can be used to increase diversity with an early number of exploration iterations.

The second contribution is directly related to the application domain. This work uses recipes as an application domain to apply the concept of dynamic critique in CRS. In the first user study, the main task for participants is to create a weekly meal plan using different variations of *DiversityBite*. The results show that participants created a diverse meal plan while exploring recipes using dynamic critique compared to non-critique-based recommenders. This suggests that diversity-focused critique enabled participants to navigate through the search space of diverse recipes. Therefore, the dynamic critique can be used as a step to increase diet diversification.

The last contribution is also related to the domain of recipes. This dissertation presented a user study that compares different forms of recipe representation for *DiversityBite*. The results show that different representation affects the diversity of the created meal plan by participants. Exploration analysis shows that participants tend to favor one set of critiques over another depending on the critique types i.e flavor and nutrition. Demographic analysis shows that there are differences among gender groups in the diversity of created meal plans deepening on recipe representation.

5.3 Limitations

To increase the diversity of recommendation through the use of critique during exploration is an open research topic. This dissertation addresses this topic by introducing a computational approach to generate critiques, in which, if a critique is selected it increases the diversity in subsequent iterations. However, there are some limitations to the present research.

First, recipe representation using ingredients resulted in sparse vectors. While using FOODON ontology reduced the number of ingredients from 11,113 ingredients to 3,807 the average number of ingredients per recipe was 11 meaning that the recipes' vectors were still sparse. This in return reduced the power of the similarity measure by reducing the contribution of rare ingredients on the similarity measure. One suggestion to overcome this problem, is to use a hierarchical ontology to map infrequent ingredients into clusters in which each cluster contains a set of similar ingredients. For example, rare spices can be placed in one cluster. This in return will not only provide a richer representation of recipes but also provides the ability to describe recipes in a more abstraction level. Another recipe representation proposed by Grace et. al. [75] in which recipes were represented as a condense vector of learnt embedding. The learnt embedding was built using the ingredient co-occurrence matrix similar to the co-occurrence matrix built for word embedding in Glove [76]. The learnt embedding resulted in placing similar ingredients closer to each other in the learnt embedding space.

Second, this work treats all users similarly in terms of their capacity for diversity. Previous research highlights the importance of personalized diversity such as [77], in which the level of diversity depends on the user's profile. The idea behind personalizing diversity is that diversity should be aligned with the user's interest, the more variety in user preferences the higher the capacity for diversity. Modeling personalized diversity requires more data that can't be obtained through running user studies. To address this, we allow participants the option of going back to the previous exploration in case the results are too diverse and happen to lie outside the user preference or too narrow and the participant would like more diversity.

Another limitation to this work is related to data availability, in the literature section, several datasets were listed that are used in the recipe domain. Even though there are a variety of datasets found in the literature, most of the datasets were created to evaluate recommender systems rather than support user exploration. For example, the number of datasets that contain user-rating data exceeds the number of datasets that contain meta-data describing the content of recipes. Due to the nature of critique by relying on features related to the item this work is based on a dataset that has the most features that could be used to support critique. The data availability also proposes challenges in conducting offline experiments in critique-based recommendation. Since most of the dataset with real user interaction do not provide critique selection, simulating users' selection in offline experiments will not be based on real users' selection. This work adopts a random approach to select a critique and mitigate this issue by simulating a prolonged user interaction. From a dynamic critique perspective, the lack of the data proposes challenges in evaluating the quality of the critique selection and whether it aligns with users' profile. However, the results of the user studies were close to the results obtained from the offline experiment which are inline with the results of the proposed offline approach.

Finally, this work is not focused on increasing the accuracy of the recommended recipes. However, the balance should be kept between accuracy and diversity. While the findings of this work are diversity-related, this work keeps this balance by introducing a similarity component to make sure the recommended recipes are not beyond the user's interest. Although there are limitations, this dissertation has introduced and evaluated a new dynamic critique approach that is diversity-driven.

5.4 Future Research Directions

This dissertation has introduced a dynamic critique approach to incorporate diversity. There are three distinct paths for future research that will be discussed in this section. The first section discusses the introduction of personalized diversity. The second section discusses incorporating nutritional information in the recommendation while exploring. The third section discusses research to generalize the diversity-focused critique onto different fields.

5.4.1 Personalizing Diversity

This dissertation focuses on introducing diversity by finding critique that increases diversity during user interaction. This work can be extended by introducing a personalized diversity that takes into account users' propensities in terms of diversity. Personalized diversity has been discussed in the literature [77, 78, 79, 77, 80] however incorporating personalized diversity in critique-based recommenders is still an open research problem. For example, users with a high propensity for diversity can be recommended a variety of recipes from different cuisines compared to a user with a low propensity for diversity.

5.4.2 Health-focused Dynamic Critique

Diet diversification has short- and long-term positive outcomes as mentioned in the literature section. Increasing diversity in recipe recommenders is one step toward healthy exploration. This work acknowledges the importance of making diversity bounded by healthy constraints. In other words, the proposed dynamic critique may increase diet diversification but it does not focus on the need to lead to healthy choices. For example, introducing diversity bounded by health constraints will drop critiques that may introduce unhealthy options such as recipes with more saturated fat.

5.4.3 Application on Different Fields

The focus of this dissertation is on increasing diversity in the recipes domain. The presented approach can be expanded to other domains that benefit from recommendation diversity. Music recommendation is one of the application domains that benefit from introducing diversity. In [81], the authors highlighted the importance of introducing diversity in music recommendation which increases the confidence of users' choices. The approach presented in the study uses visualization techniques to increase diversity, other studies show that users prefer diverse music recommendations such as study [82]. Songs are rich in features that can be used in critique such as genre, release year, artist, band, etc. While most music recommender algorithms use acoustic features for the recommendation, the use of critique will enable the user to explore songs that depend on features the user can relate to.

5.5 Conclusion

This dissertation explored the role of using dynamic critique in critique-based conversational recommender systems to support diet diversification. Diversity of recommendation is achieved by using diversity-driven critiques that are generated during the exploration process. While diversity may reduce the relevance of recommendation, the conversational recommender kept the balance between relevance and diversity by keeping the user in the loop through applying critique.

The aspects of generating dynamic critique to increase diversity in recipe recommender presented in this dissertation are used as an aspiration to design and develop DiversityBite, which is an interactive critique-based conversational recommender system designed to support users in finding diverse recipes. DiversityBite utilizes several features of recipes to provide recommendations, these features are ingredients, nutrition, and flavor with the ability to adopt new features if emerged. Exploration in DiversityBite consists of three main steps: recommend, review, and revise. The recommended step involves finding recipes that match the user's preference and generating critiques for each recipe. In the review step, the user navigates through recommended recipes and critiques. Finally, in the revise step, the user selects a critique which serves as a feedback input to the recommended step.

The simulation study and the user studies presented in this dissertation address the research questions outlined in the first chapter. The simulation study was adopted to understand the diversity of the recommended recipes along with the effectiveness of the presented dynamic critique approach. Depending on the outcomes of the simulation study, the two user studies are designed to allow participants to create a weekly meal plan. The first user study compares the proposed dynamic critique approach with static critique, along with similarity-based recommenders as a baseline. The comparison was focused on the diversity aspect of the recommendation in each variation. The second user study addresses the effect of using different features of recipes on diversity during the exploration. The focus of the study is to find any differences in the diversity while using different features of recipes.

This dissertation shows that using diversity-focused dynamic critique increases the diversity of the recommended items. This has positive outcomes in terms of enabling users to explore more options in a reasonable amount of time. The user studies in this work show that users were able to explore more recipes and compile a diverse meal plan compared to non-critique-based recommender. This work also shows that specific dimension of diversity can be increased by carefully choosing a specific underlying representation during the critique generation process.

REFERENCES

[1] K. McCarthy, J. Reilly, L. McGinty, and B. Smyth, “An analysis of critique diversity in case-based recommendation.,” in *FLAIRS Conference*, pp. 123–128, 2005.

[2] L. Chen and P. Pu, “Critiquing-based recommenders: survey and emerging trends,” *User Modeling and User-Adapted Interaction*, vol. 22, no. 1-2, pp. 125–150, 2012.

[3] W. Min, S. Jiang, L. Liu, Y. Rui, and R. Jain, “A survey on food computing,” *ACM Computing Surveys (CSUR)*, vol. 52, no. 5, pp. 1–36, 2019.

[4] L. S. Drescher, S. Thiele, and G. B. Mensink, “A new index to measure healthy food diversity better reflects a healthy diet than traditional measures,” *The Journal of nutrition*, vol. 137, no. 3, pp. 647–651, 2007.

[5] B. E. Kahn and B. Wansink, “The influence of assortment structure on perceived variety and consumption quantities,” *Journal of consumer research*, vol. 30, no. 4, pp. 519–533, 2004.

[6] H. D. Braude, *Intuition in medicine: A philosophical defense of clinical reasoning*. University of Chicago Press, 2012.

[7] M. Barker and J. Swift, “The application of psychological theory to nutrition behaviour change: Workshop on âchanging nutrition behaviour to improve maternal and fetal healthâ,” *Proceedings of the Nutrition Society*, vol. 68, no. 2, pp. 205–209, 2009.

[8] K. Grace, M. Maher, D. Wilson, and N. Najjar, “Personalised specific curiosity for computational design systems,” 2017.

[9] J. Freyne and S. Berkovsky, “Intelligent food planning: personalized recipe recommendation,” in *Proceedings of the 15th international conference on Intelligent user interfaces*, pp. 321–324, 2010.

[10] E. Pariser, *The filter bubble: How the new personalized web is changing what we read and how we think*. Penguin, 2011.

[11] D. Jannach, A. Manzoor, W. Cai, and L. Chen, “A survey on conversational recommender systems,” *arXiv preprint arXiv:2004.00646*, 2020.

[12] W. Wang and I. Benbasat, “Research noteâa contingency approach to investigating the effects of user-system interaction modes of online decision aids,” *Information Systems Research*, vol. 24, no. 3, pp. 861–876, 2013.

[13] B. Smyth and L. McGinty, “An analysis of feedback strategies in conversational recommenders,” in *the Fourteenth Irish Artificial Intelligence and Cognitive Science Conference (AICS 2003)*, Citeseer, 2003.

[14] M. Kaminskas and D. Bridge, "Diversity, serendipity, novelty, and coverage: a survey and empirical analysis of beyond-accuracy objectives in recommender systems," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 7, no. 1, pp. 1–42, 2016.

[15] A. Barraza-Urbina, B. Heitmann, C. Hayes, and A. Carrillo-Ramos, "Xplodiv: An exploitation-exploration aware diversification approach for recommender systems," in *The twenty-eighth international flairs conference*, 2015.

[16] P. Wärnestål, "User evaluation of a conversational recommender system," in *Proceedings of the 4th Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, Citeseer, 2005.

[17] Y. Sun and Y. Zhang, "Conversational recommender system," in *The 41st international acm sigir conference on research & development in information retrieval*, pp. 235–244, 2018.

[18] P. Pu and L. Chen, "Integrating tradeoff support in product search tools for e-commerce sites," in *Proceedings of the 6th ACM conference on Electronic commerce*, pp. 269–278, 2005.

[19] R. D. Burke, K. J. Hammond, and B. C. Young, "Knowledge-based navigation of complex information spaces," in *Proceedings of the national conference on artificial intelligence*, vol. 462, p. 468, 1996.

[20] P. Pu, L. Chen, and R. Hu, "Evaluating recommender systems from the userâs perspective: survey of the state of the art," *User Modeling and User-Adapted Interaction*, vol. 22, no. 4-5, pp. 317–355, 2012.

[21] J. Reilly, K. McCarthy, L. McGinty, and B. Smyth, "Dynamic critiquing," in *European Conference on Case-Based Reasoning*, pp. 763–777, Springer, 2004.

[22] K. McCarthy, J. Reilly, L. McGinty, and B. Smyth, "On the dynamic generation of compound critiques in conversational recommender systems," in *International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems*, pp. 176–184, Springer, 2004.

[23] J. Carbonell and J. Goldstein, "The use of mmr, diversity-based reranking for reordering documents and producing summaries," in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 335–336, 1998.

[24] H. Markowitz, "Portfolio selection," 1959.

[25] M. Madary and T. K. Metzinger, "Real virtuality: a code of ethical conduct. recommendations for good scientific practice and the consumers of vr-technology," *Frontiers in Robotics and AI*, vol. 3, p. 3, 2016.

[26] K. Bradley and B. Smyth, “Improving recommendation diversity,” in *Proceedings of the Twelfth Irish Conference on Artificial Intelligence and Cognitive Science, Maynooth, Ireland*, pp. 85–94, Citeseer, 2001.

[27] D. M. Fleder and K. Hosanagar, “Recommender systems and their impact on sales diversity,” in *Proceedings of the 8th ACM conference on Electronic commerce*, pp. 192–199, 2007.

[28] B. Smyth and P. McClave, “Similarity vs. diversity,” in *International conference on case-based reasoning*, pp. 347–361, Springer, 2001.

[29] C.-N. Ziegler, S. M. McNee, J. A. Konstan, and G. Lausen, “Improving recommendation lists through topic diversification,” in *Proceedings of the 14th international conference on World Wide Web*, pp. 22–32, 2005.

[30] S. Vargas and P. Castells, “Rank and relevance in novelty and diversity metrics for recommender systems,” in *Proceedings of the fifth ACM conference on Recommender systems*, pp. 109–116, 2011.

[31] M. D. Ekstrand, F. M. Harper, M. C. Willemsen, and J. A. Konstan, “User perception of differences in recommender algorithms,” in *Proceedings of the 8th ACM Conference on Recommender systems*, pp. 161–168, 2014.

[32] J. P. Kelly and D. Bridge, “Enhancing the diversity of conversational collaborative recommendations: a comparison,” *Artificial Intelligence Review*, vol. 25, no. 1-2, pp. 79–95, 2006.

[33] M. T. Ribeiro, A. Lacerda, A. Veloso, and N. Ziviani, “Pareto-efficient hybridization for multi-objective recommender systems,” in *Proceedings of the sixth ACM conference on Recommender systems*, pp. 19–26, 2012.

[34] S. Vargas, L. Baltrunas, A. Karatzoglou, and P. Castells, “Coverage, redundancy and size-awareness in genre diversity for recommender systems,” in *Proceedings of the 8th ACM Conference on Recommender systems*, pp. 209–216, 2014.

[35] V. Dang and W. B. Croft, “Diversity by proportionality: an election-based approach to search result diversification,” in *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pp. 65–74, 2012.

[36] L. McGinty and B. Smyth, “On the role of diversity in conversational recommender systems,” in *International Conference on Case-Based Reasoning*, pp. 276–290, Springer, 2003.

[37] C. Anderson, “A survey of food recommenders,” *arXiv preprint arXiv:1809.02862*, 2018.

[38] F. Ricci, L. Rokach, and B. Shapira, “Introduction to recommender systems handbook,” in *Recommender systems handbook*, pp. 1–35, Springer, 2011.

[39] C. Trattner and D. Elsweiler, "Food recommender systems: important contributions, challenges and future research directions," *arXiv preprint arXiv:1711.02760*, 2017.

[40] J. Freyne, S. Berkovsky, and G. Smith, "Recipe recommendation: accuracy and reasoning," in *International conference on user modeling, adaptation, and personalization*, pp. 99–110, Springer, 2011.

[41] M. Harvey, B. Ludwig, and D. Elsweiler, "You are what you eat: Learning user tastes for rating prediction," in *International symposium on string processing and information retrieval*, pp. 153–164, Springer, 2013.

[42] T. Kusmierczyk and K. Nørvåg, "Online food recipe title semantics: Combining nutrient facts and topics," in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pp. 2013–2016, 2016.

[43] J. Jermsurawong and N. Habash, "Predicting the structure of cooking recipes," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 781–786, 2015.

[44] M. Harvey and D. Elsweiler, "Automated recommendation of healthy, personalised meal plans," in *Proceedings of the 9th ACM Conference on Recommender Systems*, pp. 327–328, 2015.

[45] M. Ge, M. Elahi, I. Fernaández-Tobías, F. Ricci, and D. Massimo, "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.

[46] C. Trattner and D. Elsweiler, "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.

[47] T. Griffiths, "Gibbs sampling in the generative model of latent dirichlet allocation (2002)," *URL http://citeseerx. ist. psu. edu/viewdoc/summary*, 2002.

[48] Y. Hu, Y. Koren, and C. Volinsky, "Collaborative filtering for implicit feedback datasets," in *2008 Eighth IEEE International Conference on Data Mining*, pp. 263–272, Ieee, 2008.

[49] M. Rokicki, E. Herder, T. Kuśmierczyk, and C. Trattner, "Plate and prejudice: Gender differences in online cooking," in *Proceedings of the 2016 conference on user modeling adaptation and personalization*, pp. 207–215, 2016.

[50] T. Kusmierczyk, C. Trattner, and K. Nørvåg, "Temporal patterns in online food innovation," in *Proceedings of the 24th international conference on world wide web*, pp. 1345–1350, 2015.

[51] C. Trattner, M. Rokicki, and E. Herder, "On the relations between cooking interests, hobbies and nutritional values of online recipes: Implications for health-aware recipe recommender systems," in *Adjunct publication of the 25th conference on user modeling, adaptation and personalization*, pp. 59–64, 2017.

[52] H. Cheng, M. Rokicki, and E. Herder, "The influence of city size on dietary choices and food recommendation," in *Proceedings of the 25th conference on user modeling, adaptation and personalization*, pp. 359–360, 2017.

[53] M. De Choudhury, S. Sharma, and E. Kiciman, "Characterizing dietary choices, nutrition, and language in food deserts via social media," in *Proceedings of the 19th acm conference on computer-supported cooperative work & social computing*, pp. 1157–1170, 2016.

[54] Y.-Y. Ahn, S. E. Ahnert, J. P. Bagrow, and A.-L. Barabási, "Flavor network and the principles of food pairing," *Scientific reports*, vol. 1, no. 1, pp. 1–7, 2011.

[55] M. Rokicki, E. Herder, and C. Trattner, "How editorial, temporal and social biases affect online food popularity and appreciation," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 11, 2017.

[56] M. Ge, F. Ricci, and D. Massimo, "Health-aware food recommender system," in *Proceedings of the 9th ACM Conference on Recommender Systems*, pp. 333–334, 2015.

[57] D. Elsweiler, M. Harvey, B. Ludwig, and A. Said, "Bringing the" healthy" into food recommenders.," in *DMRS*, pp. 33–36, 2015.

[58] Y. van Pinxteren, G. Geleijnse, and P. Kamsteeg, "Deriving a recipe similarity measure for recommending healthful meals," in *Proceedings of the 16th international conference on Intelligent user interfaces*, pp. 105–114, 2011.

[59] C.-Y. Teng, Y.-R. Lin, and L. A. Adamic, "Recipe recommendation using ingredient networks," in *Proceedings of the 4th Annual ACM Web Science Conference*, pp. 298–307, 2012.

[60] K.-J. Kim and C.-H. Chung, "Tell me what you eat, and i will tell you where you come from: A data science approach for global recipe data on the web," *IEEE Access*, vol. 4, pp. 8199–8211, 2016.

[61] J. Chen and C.-W. Ngo, "Deep-based ingredient recognition for cooking recipe retrieval," in *Proceedings of the 24th ACM international conference on Multimedia*, pp. 32–41, 2016.

[62] S. Sajadmanesh, S. Jafarzadeh, S. A. Ossia, H. R. Rabiee, H. Haddadi, Y. Mejova, M. Musolesi, E. D. Cristofaro, and G. Stringhini, "Kissing cuisines: Exploring worldwide culinary habits on the web," in *Proceedings of the 26th international conference on world wide web companion*, pp. 1013–1021, 2017.

[63] J. Chen, L. Pang, and C.-W. Ngo, "Cross-modal recipe retrieval: How to cook this dish?," in *International Conference on Multimedia Modeling*, pp. 588–600, Springer, 2017.

[64] A. Salvador, N. Hynes, Y. Aytar, J. Marin, F. Ofli, I. Weber, and A. Torralba, "Learning cross-modal embeddings for cooking recipes and food images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3020–3028, 2017.

[65] W. Min, S. Jiang, J. Sang, H. Wang, X. Liu, and L. Herranz, "Being a supercook: Joint food attributes and multimodal content modeling for recipe retrieval and exploration," *IEEE Transactions on Multimedia*, vol. 19, no. 5, pp. 1100–1113, 2016.

[66] W. Min, B.-K. Bao, S. Mei, Y. Zhu, Y. Rui, and S. Jiang, "You are what you eat: Exploring rich recipe information for cross-region food analysis," *IEEE Transactions on Multimedia*, vol. 20, no. 4, pp. 950–964, 2017.

[67] M. Rokicki, C. Trattner, and E. Herder, "The impact of recipe features, social cues and demographics on estimating the healthiness of online recipes," in *Twelfth International AAAI Conference on Web and Social Media*, 2018.

[68] F. Abbas, N. Najjar, and D. Wilson, "The bites eclectic: Critique-based conversational recommendation for diversity-focused meal planning," in *International Conference on Case-Based Reasoning*, pp. 1–16, Springer, 2021.

[69] F. G. Abbas, N. Najjar, and D. Wilson, "Critique generation to increase diversity in conversational recipe recommender system," in *The International FLAIRS Conference Proceedings*, vol. 34, 2021.

[70] F. Abbas, N. Najjar, and D. Wilson, "Increasing diversity through dynamic critique in conversational recipe recommendations," in *Proceedings of the 13th International Workshop on Multimedia for Cooking and Eating Activities*, pp. 9–16, 2021.

[71] E. J. Griffiths, D. M. Dooley, P. L. Buttigieg, R. Hoehndorf, F. S. Brinkman, and W. W. Hsiao, "Foodon: A global farm-to-fork food ontology.," in *ICBO/BioCreative*, pp. 1–2, 2016.

[72] J.-C. Chen, "Dijkstraâs shortest path algorithm," *Journal of formalized mathematics*, vol. 15, no. 9, pp. 237–247, 2003.

[73] J. Goldberger, G. E. Hinton, S. Roweis, and R. R. Salakhutdinov, "Neighbourhood components analysis," *Advances in neural information processing systems*, vol. 17, 2004.

[74] R. Hu and P. Pu, "Helping users perceive recommendation diversity.," in *DiveRS@ RecSys*, pp. 43–50, 2011.

[75] K. Grace, M. L. Maher, N. Davis, and O. Eltayeby, "Surprise walks: Encouraging users towards novel concepts with sequential suggestions," in *Proceedings of the 9th International Conference on Computational Creativity (ICCC 2018). Association of Computational Creativity*, 2018.

[76] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

[77] F. Eskandanian, B. Mobasher, and R. Burke, "A clustering approach for personalizing diversity in collaborative recommender systems," in *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*, pp. 280–284, 2017.

[78] Y. Wang, X. Zhang, Z. Liu, Z. Dong, X. Feng, R. Tang, and X. He, "Personalized re-ranking for improving diversity in live recommender systems," *arXiv preprint arXiv:2004.06390*, 2020.

[79] T. Di Noia, V. C. Ostuni, J. Rosati, P. Tomeo, and E. Di Sciascio, "An analysis of users' propensity toward diversity in recommendations," in *Proceedings of the 8th ACM Conference on Recommender Systems*, pp. 285–288, 2014.

[80] Y. Shi, X. Zhao, J. Wang, M. Larson, and A. Hanjalic, "Adaptive diversification of recommendation results via latent factor portfolio," in *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pp. 175–184, 2012.

[81] Y. Jin, N. Tintarev, and K. Verbert, "Effects of individual traits on diversity-aware music recommender user interfaces," in *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization*, pp. 291–299, 2018.

[82] M. Onori, A. Micarelli, and G. Sansonetti, "A comparative analysis of personality-based music recommender systems.," in *Empire@ RecSys*, pp. 55–59, 2016.