Food Recommender Systems: Important Contributions, Challenges and Future Research Directions

Christoph Trattner
University of Bergen
132 PUBLICATIONS **893** CITATIONS

David Elsweiler
Universität Regensburg
86 PUBLICATIONS **710** CITATIONS

Food Recommender Systems
Important Contributions, Challenges and Future Research Directions

Christoph Trattner
MODUL University Vienna
christoph.trattner@modul.ac.at

David Elsweiler
University of Regensburg
David.Elsweiler@sprachlit.uni-regensburg.de

The recommendation of food items is important for many reasons. Attaining cooking inspiration via digital sources is becoming evermore popular; as are systems, which recommend other types of food, such as meals in restaurants or products in supermarkets. Researchers have been studying these kinds of systems for many years, suggesting not only that can they be a means to help people find food they might want to eat, but also help them nourish themselves more healthily. This paper provides a summary of the state-of-the-art of so-called food recommender systems, highlighting both seminal and most recent approaches to the problem, as well as important specializations, such as food recommendation systems for groups of users or systems which promote healthy eating. We moreover discuss the diverse challenges involved in designing recsys for food, summarise the lessons learned from past research and outline what we believe to be important future directions and open questions for the field. In providing these contributions we hope to provide a useful resource for researchers and practitioners alike.

Introduction

Online recommender systems have proved to be useful in diverse situations by empowering the user to overcome the information overload problem, assisting with the decision making process and serving as a means to change user behavior (Ricci, Rokach & Shapira, 2011). One domain, which has historically received comparatively little attention, however, especially when compared to areas relating to leisure and entertainment, is the recommendation of food items. This is surprising given the importance of food for human sustenance, the range of options available, the fact that making food choices is particularly challenging (Scheibehenne, Greifeneder & Todd, 2010), and the high personal and societal costs of poor choices. Worldwide, lifestyle- and diet-related illnesses, such as obesity and diabetes, account for 60% of total deaths (Beaglehole, 2016). Both are conditions, which can be prevented and sometimes even reversed by appropriate dietary choices (Ornish et al., 1990).

As such, health-aware food recommender systems are often mooted as an important part of the solution to encourage healthier nutritional choices (Freyne & Berkovsky, 2010; Freyne, Berkovsky & Smith, 2011; Harvey, Ludwig & Elsweiler, 2012, 2013).

There are many reasons, however, which make food recommendation challenging, not only in terms of encouraging healthy behaviour, but also in predicting what people would like to eat because this is complex, multi-faceted, culturally determined, not to mention context-dependent. Moreover, when developing food recommendation systems, there are additional issues for practitioners and researchers to consider, which do not arise in other recommendation domains. These include that users may have complex, constrained needs, such as allergies or life-style preferences, such as the desire to eat only vegan or vegetarian food. In such cases, standard approaches work poorly and adequate data sources to filter recipes are not freely available. Other challenges include food items may have multiple names, ingredients can be prepared in different ways and unlike domains where products or media are recommended, it is not always clear if a recommended item can be prepared or consumed due to the potential for poor availability of ingredients, cooking knowledge or equipment.

This paper makes two primary contributions. Firstly, we provide a summary of the state-of-the-art in food recommender systems, highlighting both seminal and most recent approaches to the problem, as well as important specializations, such as food recommendation systems for groups of users or systems which promote healthy eating. We examine which algorithms have been used in the food domain, how systems are typically evaluated, and the resources available to those interested in building or studying recommender systems in practice. In a second contribution we discuss the diverse challenges involved, as well as a summary of the lessons learned from past research and an outline of important future directions and open questions. In providing these contributions we hope to provide a useful resource for researchers and practitioners alike.

Developed Approaches

Despite food recommendation being a comparatively understudied problem in the research community, a decent body of literature exists. Table 1 provides a list of important re-

Table 1
Overview of different types of recommender systems strategies developed for recommending food (recipes, meal plans, groceries and menus) to people sorted in chronological order by publication date.

Author(s)	**Algorithm(s)**	**Person-alized**	**RecSys Type(s)**	**Feedback**	**Context/ Content Feature(s)**	**Dietary Constrains**	**Target**	**Dataset**
(Elsweiler, Trattner & Harvey, 2017)	Logistic Random Forrest Naive Bayes	no	Recipes	Ratings Binary	Title Image Ingredients Nutrition Pop. & Appr	no	Single User	Allrecipes
(Trattner & Elsweiler, 2017)	LDA WRMF AR SLIM BPR MostPop User- ItemKNN	yes/no	Recipes Meal Plans	Bookmarks Ratings Comments	WHO-FSA health score	no	Single User	Allrecipes
(Cheng, Rokicki & Herder, 2017)	BPR MostPop	yes/no	Recipes	Ratings	City Size	no	Single User	Kochbar
(Yang et al., 2017)	Learning to Rank	yes	Recipes	Binary	Image Embeddings	yes	Single User	Yummly
(Rokicki, Herder, Kuśmierczyk & Trattner, 2016)	UserKNN MostPop	yes/no	Recipes	Ratings	Gender	no	Single User	Kochbar
(Ge, Elahi, Fernaández-Tobías, Ricci & Massimo, 2015)	MF CB	yes	Recipes	Ratings Tags	Tags	no	Single User	Wellbeing Diet Book
(Elsweiler & Harvey, 2015)	SVD-Hybrid	yes	Meal Plans (Set of recipes)	Ratings	Ingredients	yes	Single User	Quizine
(Sano, Machino, Yada & Suzuki, 2015)	UserKNN SVD Hybrid NL-PCA	yes	Groceries	Purchases	Food Categories	no	Single User	Grocery store data
(Trevisiol, Chiarandini & Baeza-Yates, 2014)	UserKNN CB	yes	Menus (Set of dishes)	Binary	Text Sentiment	no	Single User	Yelp
(Elahi, Ge, Ricci, Massimo & Berkovsky, 2014)	MF	yes	Recipes	Ratings Tags	tags	no	Group of Users	Wellbeing Diet Book
(Harvey et al., 2013)	CB, CF Logistic Reg. SVD-Hybrid	yes	Recipes	Ratings	Ingredients etc.	no	Single User	Quizine
(Teng, Lin & Adamic, 2012)	SVM	no	Recipes	Ratings	Ingredients Nutrition Cook effort Cook methods	no	Single User	Allrecipes
(Kuo, Li, Shan & Lee, 2012)	Graph-based CB	yes	Menus (Set of recipes)	Tags	Ingredients	no	Single User	Food
(El-Dosuky, Rashad, Hamza & El-Bassiouny, 2012)	CB KB	yes	Food items	Query	tags	no	Single User	USDA
(Freyne, Berkovsky, Baghaei, Kimani & Smith, 2011)	CF	yes	Meal plans (Set of recipes)	Ratings	-	no	Single User	Wellbeing Diet Book
(Ueta, Iwakami & Ito, 2011)	KB	yes	Recipes	Query	tags	no	Single User	Cookpad
(van Pinxteren, Geleijnse & Kamsteeg, 2011)	CB	yes	Recipes	Cooked recipes	Recipe content features	no	Single User	Smulweb
(Freyne & Berkovsky, 2010)	UserKNN CB Hybrid	yes	Recipes	Ratings	Ingredients	no	Single User	Wellbeing Diet Book
(Berkovsky & Freyne, 2010)	UserKNN GroupKNN Hybrid	yes	Recipes	Ratings	-	no	Group of Users	Wellbeing Diet Book
(Aberg, 2006)	CF	yes	Meal Plans (Set of recipes)	Ratings	-	yes	Single User	Unknown
(Khan & Hoffmann, 2003)	CBR	yes	Meal Plans	Query	Nutrition Content	yes	Single User	Unknown
(Mankoff, Hsieh, Hung, Lee & Nitao, 2002)	CB	yes	Groceries	Purchases	Food groups	no	Single User	Grocery store data
(Lawrence, Almasi, Kotlyar, Viveros & Duri, 2001)	AR CF CB	yes	Groceries	Purchases	Product class	no	Single User	Grocery store data
(Hinrichs & Kolodner, 1991)	CBR	yes	Meal Plans	Query	Content	yes	Single User Group of Users	Unknown
(Hammond, 1986)	CBR	yes	Single New Recipe	Query	-	yes	Single User	Unknown

search contributions relating to food recommendation. We list 25 popular, highly cited, recent and relevant papers in chronological order, selected using our experience in the domain in combination with bibliographic tools, such as Google scholar[1], to identify the most relevant for the targeted readership. Special care was taken to identify work relating to different types of food item. As such, the papers and articles cited relate to the recommendation of recipes, meal plans, groceries and menus. Although the problem of recommending restaurants to people, e.g. (Park, Park & Cho, 2008), is related, especially when the meals served there are taken into consideration, we focus here on research relating to systems directly recommending the food items themselves.

The columns in Table 1 relate to dimensions that we believe characterize the nature of different contributions in the area. *Algorithm* defines the various algorithmic approaches that have been tested in the food domain ranging from content based approaches, to collaborative filtering, to machine learning classifiers, some of which involve personalization (*Personalized*). *Recommended Items* describes the food item involved; *Feedback* describes the means by which the system is informed on user preferences and the suitability of any recommendation provided; *Context* provides the context dimension(s) utilised if applicable; *dietary constraint* informs on whether nutrition was considered; *Target* details who the end user(s) of the system was(were); and finally *Dataset* details the proprietary or open dataset utilized. The remainder of the paper uses Table 1 as a structural basis.

In this section, we explain the approaches that have been taken in the literature to implement food item recommenders. In the literature the most prominent form of food recommender system provides single item recommendations mostly in the form of recipes.

We structure the section around the approaches employed, summarizing content-based, collaborative filtering and hybrid approaches. We continue to show how context information is important and how this has been utilized in practice. Next, we broaden our focus to particular scenarios, which have been addressed, firstly looking at group-recommendations before reviewing research on food recommenders for healthy nutrition.

Reflecting the literature as a whole, the majority of section details work recommending recipes to end users. That being said, there are other kinds of food items, which have been studied, albeit to a lesser extent. This is reflected in Table 1. As datasets are becoming more readily available (see Section ‘Implementation Resources’), we expect interest in other food items to increase. The work published to date has largely employed the same standard approaches as have been applied to recipes. For example, content-based, collaborative filtering and hybrid approaches have been applied to restaurant review data to recommend menus (Trevisiol et al., 2014) and online shopping data to recommend groceries (Lawrence et al., 2001; Mankoff et al., 2002; Sano et al., 2015).

Content-Based Methods (CB)

Content-based approaches have been used as a means to tailor recommendations to the user’s individual tastes. Freyne and Berkovsky, for example, made recommendations by breaking recipes down into individual ingredients and scoring based on the ingredients contained within recipes, which users had rated positively (Freyne & Berkovsky, 2010; Freyne, Berkovsky & Smith, 2011). That is, if tomatoes had been present in recipes a user had reported liking, further recipes containing tomatoes would be predicted to also be liked by the user. Later work progressed this approach by not only accounting for positive ingredient biases, but also negatively weighting recipes based on contained ingredients featuring in recipes the user reported disliking (Harvey et al., 2013).

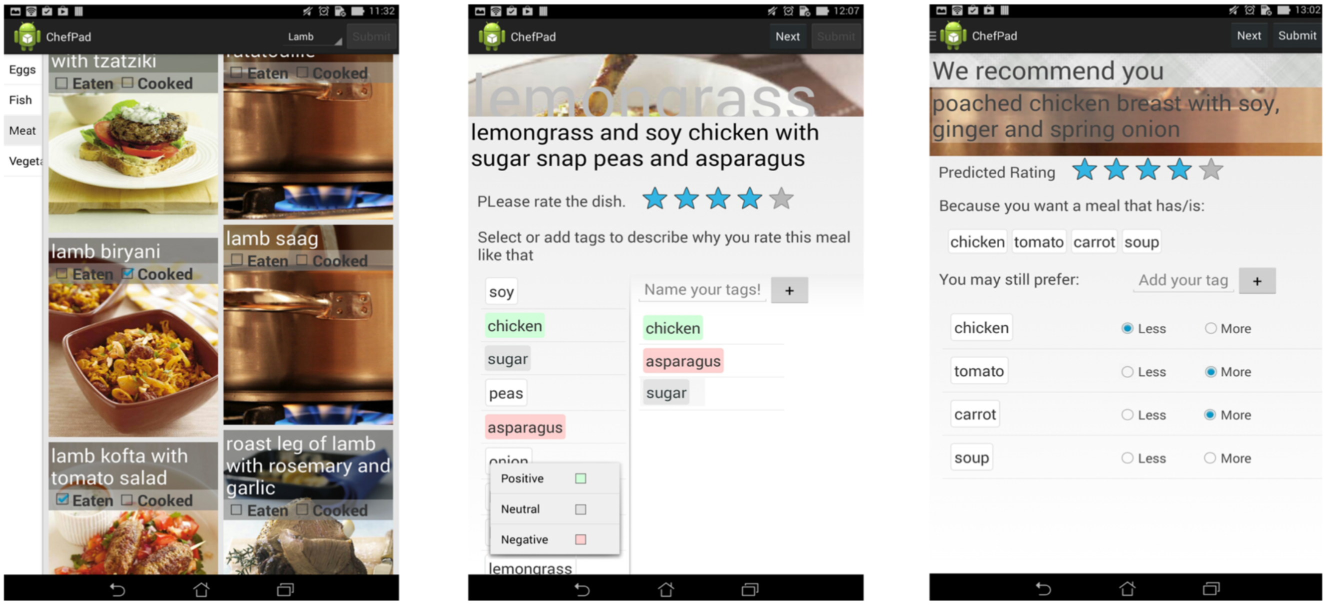
Teng et al. (2012) proposed the use of complement and substitution networks as a means to generate accurate predictions. Complement networks of ingredients are constructed via co-occurrence of the the same ingredients in the same recipes, while substitute networks are derived from user-generated suggestions for modifications. Experiments show that the use of these networks can predict the user preferences significantly better than approaches that rely on for example ingredient lists as features, cooking method, style, etc.

Other content-based approaches are more applicable to food recommender systems than other domains. For example, as food decisions are often visually driven (Mormann, Navalpakkam, Koch & Rangel, 2012; Schur et al., 2009), the images associated with recipes can be exploited. Yang and colleagues have shown baseline approaches can be outperformed by algorithms designed to extrapolate important visual aspects of food images (Yang et al., 2015, 2017). In their work, Convolutional Neural Networks (CNN) provide a powerful framework for automatic feature learning. Elsweiler, Trattner and Harvey (2017) also show that automatically extracted low-level image features, such as brightness, colorfulness and sharpness can be useful for predicting user food preference.

Collaborative Filtering-Based Methods (CF)

Collaborative filtering-based recommendation methods for food reccommender systems have also been proposed and evaluated. Freyne and Berkovsky tested a nearest neighbour approach using Pearson correlation on the ratings matrix, which offered poorer performance than the content approach described above (Freyne & Berkovsky, 2010). Harvey et al. (2013) showed that SVD outperformed both the content and collaborative filtering approaches suggested in (Freyne

[1] http://scholar.google.com



□ Eaten □ Cooked
□ Eaten ☑ Cooked
☑ Eaten □ Cooked

lemongrass and soy chicken with sugar snap peas and asparagus

PLease rate the dish. ★★★★☆

Select or add tags to describe why you rate this meal like that

poached chicken breast with soy, ginger and spring onion

Predicted Rating ★★★★☆

Because you want a meal that has/is: chicken tomato carrot soup

You may still prefer:

chicken	● Less	○ More
tomato	○ Less	● More
carrot	○ Less	● More
soup	○ Less	○ More

Figure 1. Example of a mobile food recommender interface as proposed by Elahi et al. (2014) using not only ratings for preference elicitation but also tags at the same time. Taken with permission from the authors' work.

& Berkovsky, 2010). Ge, Elahi et al. (2015) propose a matrix factorization (MF) approach for food recommender systems that fuses ratings information and user supplied tags to achieve significantly better prediction accuracy than content-based and standard matrix factorization baselines. They also present a mobile interface for the approach as shown in Figure 1. These screenshots show how a finer granularity of feedback can assigned via tags, complementing the standard binary and scaled ratings typically used.

More recently, Trattner and Elsweiler (2017) tested a diverse range of collaborative filtering approaches implemented in the LibRec[2] framework using a large dataset crawled from the online recipe portal allrecipes.com. The highest performing CF approaches were Latent Dirichlet Allocation (LDA) (Griffiths, 2002) and Weighted matrix factorization (WRMF) (Hu, Koren & Volinsky, 2008). The results of their experiments are shown in Table 2 below.

Hybrid Methods (Hybrid)

Hybrid recommenders have been proposed by other scholars for the recipe recommendation task. For example, Freyne and Berkovsky (2010) combined a user-based collaborative filtering method with a content-base method. Moreover, in their follow-up work, which targeted groups of users (described in more detail in below), they employed a hybrid approach to combine three different recommender strategies in a single model, using a switching strategy. The switching was based on the ratio between the number of items rated by a user and overall number of items. Another example of a hybrid approach can be found in the work of Harvey et al. (2013) who achieved the best performance in their experiments by combining an SVD approach with user and item biases.

Context-Aware Approaches

Numerous exploratory data analyses have demonstrated that context is important in food recommendation, with gender (Rokicki et al., 2016), time (Kusmierczyk, Trattner & Nørvåg, 2015), hobbies (Trattner, Rokicki & Herder, 2017), location (Cheng et al., 2017; De Choudhury, Sharma & Kiciman, 2016; Zhu et al., 2013) and food availability (De Choudhury et al., 2016) being identified as important variables. All of these studies employed relatively simple filtering techniques to split naturalistic datasets in order to explore how recipes were rated (Freyne, Berkovsky, Baghaei et al., 2011), bookmarked (Trattner & Elsweiler, 2017), shared (Abbar, Mejova & Weber, 2015) or relate to health statistics (Trattner, Parra & Elsweiler, 2017).

Harvey et al. (2012) collected detailed context data encapsulating the ratings participants provided for their dataset, where participants could identify a broad range of factors to justify the rating assigned to a recipe as a meal to cook for dinner that day. Analyzing these with regression modeling showed that factors, such as how well the preparation steps are described, as well as the nutritional properties of the dish, the availability of ingredients and temporal factors such as day of the week have a bearing on the user's opinion of the recommendation.

[2] http://www.librec.net/

What is lacking with respect to context is an understanding of which variables are the most important and how best to account for these algorithmically (Rokicki, Herder & Trattner, 2017). Despite studying numerous factors, Harvey et al. (2013), for example, limited their algorithmic efforts to approaches with only nutritional, user and item biases.

In summary, although the problem of improving the precision of recommendations has been attended to by numerous researchers with diverse approaches, the results achieved for the recommendation of recipes to individual users, measured on standard metrics are typically poorer than in other domains. To demonstrate typical results achieved with standard approaches on recipe data, we present the results of our own experiments with standard techniques on a well-known dataset in Table 2 below. These results underline the challenge of predicting which dishes people will like and emphasize that further effort is required.

Group-Based Methods

Of course people do not always eat or make food choices alone. Often these are activities done together with friends, families or colleagues. It is well known in social psychology that the social situation within which one will eat (who is present, why they are present, and what their preferences are) influences the food choices taken (Wansink, 2006). In food recommender systems, such social contexts are addressed by group recommender systems. In this setting, a list of items is produced for a group of people rather than for an individual user. Despite the pervasiveness of shared food consumption experiences, group-based food recommender systems research has been limited, even though the earliest efforts can be traced to the early 1990's (Hinrichs & Kolodner, 1991). Berkovsky and Freyne (2010) not only studied different strategies for recommending recipes to a group of people but evaluate these methods with real users in a family scenario. In particular their work introduces four different strategies: A general strategy (which employs a most popular approach to recommend items), an aggregated model (which first combines individual user models into a single model before applying the collaborative filter), aggregated predictions strategies (which first computes CF on the individual user profiles and then combines the predicted rating) and finally a personalized strategy (which exploits a standard CF algorithm). The results show that the personlized version works the best but it was not possible to create personalized recommendations for all of the users. More recently Elahi et al. (2014) proposed a mobile interface and algorithm for food recommender system in a group-context. In addition to improving the prediction algorithm with tags, the authors use group-based preference elicitation, in which users play different roles in the food choice process. One user is designated as the group leader or cook to whom the system delivers meal recommendations based on the group utility score, which aggregates predictions using the tags and ratings of all the group members.

Health-Aware Methods

When motivating research on food recommender systems, health problems and improving nutritional habits are usually mentioned e.g. (Freyne & Berkovsky, 2010; Freyne, Berkovsky & Smith, 2011; Harvey et al., 2012, 2013). Incorporating health into the recommendation, however, has largely been a recent focus (Elsweiler, Hors-Fraile et al., 2017; Elsweiler, Ludwig, Said, Schäfer & Trattner, 2016; Schäfer et al., 2017). One means of achieving this is to incorporate nutritional aspects into the recommendation approach directly. Ge, Ricci and Massimo (2015) took this approach by accounting for calorie counts in the recommendation algorithm. They did this based on a so-called "calorie balance function" that accounts for the differences between the calories the user needs and the calories in a recipe.

Elsweiler, Harvey, Ludwig and Said (2015) refer to the trade-off for most users between recommending the user what she wants and what is nutritionally appropriate. This is a trade-off applicable for a large proportion of users (Harvey et al., 2013) and should be optimized (Elsweiler et al., 2015). The authors proposed combining to two aspects linearly as a framework for evaluating different algorithmic approaches to incorporate health in the recommendation process.

The formula (see Equation 1) illustrates the simple concept. Here, i is a given recipe, $r(\hat{i})$ is the estimated rating for recipe i, Max($r(\hat{i})$) is the maximum estimated rating over all recipes. $n(i)$ is the nutritional "error" incurred when recommending this recipe (relative to some ideal set of nutritional values). λ is a free parameter that can be set to suit the researcher/practitioner's priorities, although λ=.5 is probably preferable initially as it gives equal weighting to rating and nutrition. Note that all of these estimates are implicitly conditioned on a specific user u.

$$score(i) = \lambda \frac{r(i)}{max(r(i))} + (1 - \lambda) - 1 \times \frac{n(i)}{max(n(i))} \qquad (1)$$

Trattner and Elsweiler (2017) employed a post-filtering (see Equation 2 and 3) approach to incorporate further nutritional aspects. To post-filter items a a straightforward scoring function is applied which re-weights the scores of a recipe for a particular user based on the WHO or inverse FSA score, employing a simple multiplication. The $score_{u,i}$ in the equation stands for the score of the item i for user u and who_i, fsa_i denote the health scores for that item. The two nutrition metrics are based on widely accepted nutritional standards from The World Health Organisation (WHO) (WHO, 2003) and the United Kingdom Food Standards Agency (FSA) (FSA, 2016) (see Section 'Implementation Resources'). Their previous work had used these measures to establish the (un)healthiness of recipes from a popular Internet food portal (Trattner, Elsweiler & Howard, 2017).

$$score_{u,i,who} = score_{u,i} \cdot (who_i + 1) \tag{2}$$

$$score_{u,i,fsa} = score_{u,i} \cdot (16 - fsa_i - 4 + 1) \tag{3}$$

Table 2 describes the performance of 9 prominent recommender algorithms as implemented in the LibRec framework in Trattner and Elsweiler's experiments. The top and bottom halves of table shows the performance without and with post-filtering respectively. Full details of the experimental setup can be found in (Trattner & Elsweiler, 2017).

These experiments with post-filtering on nutritional properties show that 1) it is possible to balance and potentially optimize the trade-off between recommendation accuracy and the healthiness of recommendations, 2) some recommendation algorithms may be more (e.g., LDA and WRMF) or less suitable (e.g., MostPop and BPR) to this process.

Nevertheless the results also show that 3) while the approach shows potential benefit and future work should try to optimize the trade-off, the method by itself will not lead to healthy nutrition - at least not with the collection evaluated in this work. Despite offering a significant improvement on the standard approaches, the post-filtered results show that the best FSA and WHO scores achieved were not particularly high and are associated with extremely poor recommendation accuracy. These represent the best health values which can be achieved using an individual item recommendation approach, indicating that complementary ideas are necessary.

One such complementary approach is to combine individual recommended items for a user, such that they meet the recommended intake for that user over a longer period of time (e.g. day, week etc.). Freyne, Berkovsky, Baghaei et al. (2011) presented an interface, which allowed users to generate their own meal plans from individually recommended dishes. The recommendations were generated using the authors' hybrid approach as described above (Freyne & Berkovsky, 2010). The interface for such plans evaluated on 5000 people in Australia. To encourage variation in meal plans a decay function was applied to meals appearing regularly in plans. Users manually created plans from lists of recommendations but the lists were filtered such that only meals that could be added and still ensure plans met guidelines featured in the list of recommended items.

Harvey and Elsweiler (2015) presented a similar interface, which automated the creation of plans consisting of a combination of breakfast, lunch and dinner plus an allowance for snacks and drinks. The same authors evaluated their planning approach systematically by deriving plans from taste profiles (i.e. from users featuring in naturalistic datasets) combined with diverse personas (simulated user properties, such as height, weight, gender, age, nutritional goal (lose/gain/maintain weight) and activity level (from sedentary to highly active) (Elsweiler & Harvey, 2015). In a first step, the authors estimated ratings users with particular profiles might assign to recipes (using approaches like those described above). In a second step, following approaches from nutritional science, the recommended nutritional intake was calculated for the user persona, including the required calories, but also where these should be sourced (proteins, carbohydrates etc.). Lastly, plans were generated for a given user (persona-profile combination) by taking the top-n recommendations from the recommender system for the taste profile, splitting these into two separate sets, one for breakfasts and one for main meals and performing a full search finds *every combination* of these recipes in the sequence [breakfast, main meal, main meal] meeting the target nutritional requirements as defined above.

Using this method the authors were able to generate plans for 4025/6400 cases (63%) and at least 1 plan was generated for 58 out of the 64 (91%) user profiles and for each of the 100 personas. The authors moreover analyzed the factors, which made the development of plans challenging. When personas required a relatively high calorie intake, e.g. if the persona was tall or wanted to gain weight, the simple approach using 3 meals of fixed portions was often unable to address this properly. Similarly, profiles with little diversity in preferred ingredients were also hard to satisfy.

Substituting meals has been mooted as a further approach to influencing food choices. Elsweiler, Trattner and Harvey (2017) developed predictive models with the aim of forecasting the choices people will make. After evaluating the models for prediction accuracy using cross-validation, these were used to select recipe replacements such that users were be "nudged" towards making healthier choices. Aligning with the findings reported above, visual and nutritional features were important. A user study found that using the predictive models as the basis for recommendations, participants were significantly more likely to choose a recipe with much less fat content - the opposite of the trend that one typically sees.

Substituting ingredients within recipes has also been proposed to improve the health credentials of individual recipes healthier e.g. (Achananuparp & Weber, 2016; Teng et al., 2012). This approach has, however, yet to be evaluated properly in a nutritional context. Initial steps in this direction were taken by Kusmierczyk, Trattner and Nørvåg (2016), whose findings illustrate to what extent it may be possible to recommend a user substitute ingredients based on the user's previous recipe uploads and accounting for social-, temporal and geographic-context.

In our experience the standard approaches applied to date in the literature do not work well when dealing with specialist diets, e.g. (vegetarian, vegan or allergies)[3]. Constraint-based approaches are found surprisingly rarely in the literature.

One exception is Yang et al. (2017) who had access to

[3] We have not published our findings, but we have run several test runs with vegetarian, vegan, and gluten allergy user profiles.

Table 2
Recommender ranking accuracy sorted by nDCG and recommender accuracy post-filtered by FSA scores. The mean FSA scores of the top-5 recommended recipes are also reported along with the different average nutriens of the lists and the according FSA health labels (taken from Trattner and Elsweiler (2017)).

Algorithm	**nDCG@5**	**FSA score**	**Fat (g)**	**Sat. Fat (g)**	**Sugar (g)**	**Sodium (g)**
LDA	.0395	9.110	8.70	3.73	8.73	0.32
WRMF	.0365	9.114	9.50	3.89	8.84	0.34
AR	.0343	9.206	9.27	4.12	10.50	0.25
SLIM	.0326	8.907	9.27	3.82	7.91	0.33
BPR	.0325	9.252	8.69	3.82	7.83	0.29
MostPop	.0294	9.004	9.02	3.94	10.01	0.23
UserKNN	.024	8.985	8.96	3.73	7.98	0.31
ItemKNN	.0178	8.652	8.59	3.51	6.03	0.31
Random	.0029	8.486	8.74	3.49	5.71	0.30
FSA score post-filtered ($score_{u,i,fsa}$)						
LDA	.0321	7.323	6.51	2.42	4.03	0.29
WRMF	.0303	7.361	6.48	2.30	4.75	0.31
SLIM	.0248	7.008	6.20	2.56	2.59	0.24
AR	.0238	6.984	5.64	1.94	3.95	0.28
MostPop	.0228	7.334	5.37	2.02	2.46	0.24
BPR	.0205	6.722	6.42	2.30	4.95	0.26
UserKNN	.0168	6.722	6.88	2.73	3.33	0.33
ItemKNN	.0109	6.124	5.15	1.79	3.51	0.25
Random	.0022	4.305	1.59	0.43	1.45	0.09

the data basis to apply filters based on vegetarian, vegan and gluten-free food. A further example can be found in the nutritional science literature whereby linear programming is used to ensure Malawian children achieve the required nutritional intake recommended by experts (Ferguson, Darmon, Briend & Premachandra, 2004). As comparable datasets exist (see Section ‘Implementation Resources’ below), there is no reason why a similar approach cannot be taken to promote healthy eating patterns in other demographics.

Addressed Challenges and Problems

As should now be clear the food recommendation task brings additional challenges to those in other recsys domains. There are also standard challenges, applicable to all domains, which have been addressed, at least to some extent, in food recommender research. In this section we first relate the generic challenges and how these have been addressed or not in the food domain, before switching focus to the challenges unique to food recommendation.

User preference sources. Food recommendation research has mainly exploited explicit sources of user feedback in the form of ratings (Freyne & Berkovsky, 2010; Freyne, Berkovsky, Baghaei et al., 2011; Harvey et al., 2013), bookmarks (Trattner & Elsweiler, 2017) or shares (Abbar et al., 2015). Methods of implicit feedback have been used less often, but examples include recipe views (Wagner, Singer & Strohmaier, 2014; West, White & Horvitz, 2013) and the sentiment of reviews submitted about recipes (Trattner & Elsweiler, 2017).

User preference scarcity. To our knowledge the problems of scarcity of user feedback, illustrated by the cold-start problem and sparse matrices, has not been directly addressed in the food recommender systems literature. Rather standard solutions, which cope well, such as SVD have been applied (Harvey et al., 2013; Trattner & Elsweiler, 2017).

Offline and online evaluation of recommendations. To our knowledge, evaluation in the food recommendation domain has been almost offline. Typically, as is explained in more detail below, datasets have been created naturalistically e.g. (Trattner & Elsweiler, 2017; Trevisiol et al., 2014; Yang et al., 2017) or via user studies (Freyne & Berkovsky, 2010; Harvey et al., 2013). These datasets form the basis of offline evaluations in the form of prediction tasks. Other evaluations have taken the form of user studies, where users test interfaces in a semi-controlled (Ge, Elahi et al., 2015) or naturalistic environment (Freyne, Berkovsky, Baghaei et al., 2011). However, full-online evaluations have to our knowledge not yet been published.

Beyond accuracy. Accuracy has been the overwhelming focus of research efforts to date but nevertheless, as described above, it remains a challenge, which in the food domain, has yet to be adequately solved. Accuracy, however, is not the only important aspect to consider when recommending food. Novelty and serendipity are both properties of food recommendations, which users appreciate (Harvey et al., 2013), but to our knowledge, these are yet to be studied. Elsweiler and Harvey (2015) did acknowledge the importance of dietary diversity in their meal plan work. Moreover, the preference-healthfulness trade-off bears many similarities to traditional work on novelty and serendipity in that it involves recommending non-preferred items while minimizing the loss in precision. While preliminary research in this direction exists (Elsweiler & Harvey, 2015; Trattner & Elsweiler, 2017), there is much work to do in order to understand how to optimize this trade-off appropriately.

Recommendation visualizations and explanations. Methods of visualization and the explanation of recommendations have been, at best, implemented in a superficial way within food recommender research. Examples include the traffic light system employed by Trattner and Elsweiler (2017) and the plan meta-data provided in the demo system presented by Harvey and Elsweiler (2015). Elahi et al. (2014) provide the best example of explanations for the recommendations offered by their system as can be seen in in Figure 1. Nevertheless, only superficial evaluations of any of these systems have been published.

Other common challenges. Despite their importance generally to recommender systems, there is nothing to report from the food domain in terms of significant contributions on the issues of privacy and collaborative recommenders, scalability and distribution of collaborative recommenders or issues of robustness or attacks on food recommenders.

Challenges unique to food recommender systems. We can see from the numerous challenges yet to be addressed in the food domain, that research in this area is still preliminary. That being said, we wish to acknowledge some domain specific challenges, which have been addressed to some extent. Firstly, as Section 'Developed Approaches' shows, the challenge of tailoring standard approaches to the problem has been tackled.

There have been efforts to better process and understand the content of items to be recommended. These include normalizing ingredients and ingredient quantities (Kusmierczyk et al., 2015; Müller, Mika, Harvey & Elsweiler, 2012); understanding the role of context in user decision processes (see Section on context-aware recommender systems), and understanding which visual features are helpful in guiding these choices (Elsweiler, Trattner & Harvey, 2017; Yang et al., 2017).

With respect to health, there have been preliminary efforts to model nutritional aspects of the process (Schäfer et al., 2017), which include user requirements (Gibney, Vorster & Kok, 2002), user intake (Straßburg, 2010) and the estimation of portion sizes (Zhang et al., 2011). Other work has preprocessed recipes to establish the nutritional content either by ingredient matching (Müller et al., 2012) or by visually analyzing food images (Chokr & Elbassuoni, 2017). Finally, as we described in detail above, progress has been made in incorporating health in the recommendation process either by considering nutrition in item recommendation e.g. (Ge, Elahi et al., 2015), generating meal plans (Elsweiler & Harvey, 2015) or via algorithmic nudging (Elsweiler, Trattner & Harvey, 2017). It is unclear, however, which method works most effectively.

Implementation Resources

In this section we summarize resources that can help in the development of food recommender systems. We summarize (i) datasets typically used to study food consumption patterns and to evaluate algorithmic approaches, (ii) nutrition and health resources, available to implement health-aware recommender systems. Finally, frameworks typically employed to build these are described.

Recipe, Meal plan, Menu and Grocery Store Datasets

To date research in the food recommender systems domain typically relies on proprietary and none standardized datasets. This contrasts with domains such as movie recommender domain, where the well-known MovieLens datasets have set a standard. The following list highlights datasets usually employed when it comes to the implementation of recipe, meal plan, grocery and menu recommender systems.

Recipes. Most of the research for recommending recipes relies on Web resources, e.g., Allrecipes[4] or Food.com[5] which comprise rich item and user profiles. Although these offer an extensive basis for conducting research in that direction, most of the datasets cannot be shared as the terms of services of the sites explicitly forbid it. As such, few publicly accessible datasets comprising recipe and user profiles are available. Researchers must typically develop their own crawlers or seek a license agreement with the platform providers. The Australian government agency CSIRO'S Wellbeing Diet Book[6] has been used by Australian researchers (Freyne & Berkovsky, 2010) and connected researchers in Italy (Elahi et al., 2014), but is not readily available to other researchers. Cookpad[7] and Yummly[8] have both supported academic research by providing licensed access to recipe and profile data, and Yummly also supports broad access to restricted data via a no-cost API. One dataset has recently been made available by the Massachusetts Institute of Technology

[4] http://www.allrecipes.com

[5] http://www.food.com

[6] https://www.csiro.au/en/Research/Health/CSIRO-diets/CSIRO-Total-Wellbeing-Diet

[7] http://www.cookpad.com

[8] http://www.yummly.com

Table 3
Example of an nutrition entry for the query 'apple' in the USDA database.

Nutrition	Unit	Value per 100g
Water	g	85.56
Energy	kcal	52
Protein	g	0.26
Total lipid (fat)	g	0.17
Carbohydrate, by difference	g	13.81
Fiber, total dietary	g	2.4
Sugars, total	g	10.39

(MIT)[9] comprising of over 1 million recipes including food images and some meta-data. The dataset is limited, however, in that no user profiles or interactions are available, and as such the dataset may not be suitable for evaluating a recipe recommender system in an offline scenario. The lack of standard collections restricts the reliability and generalizability of research published to date.

Meal plans and restaurant menus. Meal plan recommender research has typically relied on the same recipe datasets as above. To our knowledge no freely available datasets containing meal plans exist. Yelp[10] has been used as a resources to build and evaluate menu recommender system algorithms. As with recipe datasets, in order to obtain the data one might need to implement a crawling framework as the terms of services of the site to date omit data sharing

Groceries. In the grocery recommender scenario, to our knowledge, only one dataset is freely available. This dataset was published by Kaggle[11] and contains 3 millions purchases of users on instacart and comprises limited meta-data (such as grocery name) in respect to the groceries bought out in a basket. Table 1 refers to some other datasets but these are not available publicly.

Nutrition & Health Resources

When it comes to the implementation of food recommender systems or algorithms it is not only beneficial to have open-data datasets comprising of user and item profiles (as discussed earlier), but also other external resources that help in building such a system. For instance, to build a health-aware recipe recommender system, it is essential to know the nutritional values of food items and to what extent these may be healthy or unhealthy To estimate nutrition, the typical approach is to map ingredients to standard databases, such as those provided by the USDA[12] (US) or the BLS[13] (Germany). As an example, Table 3 provides a partial entry for the ingredient 'apple'[14]. The example is far from being complete, as also 'Minerals', 'Vitamins', 'Lipids' and other macro nutrients can be obtained such as 'Caffeine' are also accessable in the database. One of the challenges typically involved in the matching process is the normalization of the ingredients in a recipe, as different names are often used to express the same entity, such as '100g Parmesan cheese' vs '100g of shredded Parmesan cheese'. The method of processing or cooking may additionally influence the nutritional value. Moreover, units are often not expressed using normalized units of quantity. One recipe may refer to 'one cup of water' whereas another may refer to the same item as '235ml water'. Detailed descriptions of the challenges involved can be found in (Müller et al., 2012). Standard NLP techniques such as stop-word removal, conjunction splitting, string matching, etc. can be applied to address some of these (see for example (Kusmierczyk et al., 2015)). A more practical means to extract this kind of information is though to for instance employ a Web services such as provided by SPOONACULAR[15], whose API is able to extract ingredient names and amounts in a unified way, which can in some cases be accessed for free for purposes of academic research.

Other resources to identify the nutritional properties of a meal (recipe) are provided by (Müller et al., 2012). These output the nutritional properties for a given German recipe by utilizing the BLS database. Müller employ a multi-step process, first utilising a rule-based infrastructure before a learning to rank approach to identify the most appropriate database entry for a given ingredient. The framework can be obtained from the authors without cost but a license for the BLS is required to use the software. The EDAMAM[16] Web service offers similar functionality for English and Spanish recipes. This service is a commercial product, but as with Spoonacular, can in some cases used without cost for academic purposes.

To estimate the healthiness of a meal (Trattner, Elsweiler & Howard, 2017), one may rely on standards as set by nutrition scientists. There are many of such standards for different countries and other geographical regions. The ones which have been successfully applied to the food recommender problem (see (Elsweiler, Trattner & Harvey, 2017; Trattner & Elsweiler, 2017)) are provided by the Food Standard Agency (FSA) (FSA, 2016) and the World Health Organization (WHO) (WHO, 2003). Both provide tables based on a 2000kcal diet that contain ranges of nutrients, such as for example Fat, Saturated Fat, Sugar and Sodium (see Table 4 and Table 5). The WHO guidelines account for macronutrients, such Fiber content, and so on. The FSA guidelines are typically used to derive front of package labels for meals

[9] http://im2recipe.csail.mit.edu

[10] http://www.yelp.com

[11] https://www.kaggle.com/c/instacart-market-basket-analysis

[12] https://ndb.nal.usda.gov/ndb

[13] https://www.blsdb.de

[14] https://ndb.nal.usda.gov/ndb/foods/show/2122

[15] https://market.mashape.com/spoonacular

[16] https://www.edamam.com

Table 4
FSA front of package guidelines as proposed in FSA (2016) and as, for example, used in Trattner and Elsweiler (2017).

Text	**LOW**	**MEDIUM**	**HIGH**
Color code	**Green**	**Amber**	**Red**
Fat	≤ 3.0g/100g	> 3.0g to ≤ 17.5g/100g	> 17.5g/100g or > 21g/portion
Saturates	≤ 1.5g/100g	> 1.5g to ≤ 5.0g/100g	> 5.0g/100g or > 6.0g/portion
Sugars	≤ 5.0g/100g	> 5.0g to ≤ 22.5g/100g	> 22.5g/100g or > 27g/portion
Salt	≤ 0.3g/100g	> 0.3g to ≤ 1.5g/100g	> 1.5g/100g or > 1.8g/portion

Table 5
WHO guidelines as originally proposed in WHO (2003) and adopted to recipes by Howard, Adams and White (2012) and as, for example, used in Trattner and Elsweiler (2017).

Dietary Factor	**Range (percentage of kcal per meal/recipe)**
Protein	10-15
Carbohydrates	55-75
Sugar	< 10
Fat	15-30
Saturated Fat	< 10
Fiber density (g/MJ)	> 3.0†
Sodium density (g/MJ)	< 0.2‡

†Based on 8.4 MJ/day (2,000 kcal/day) diet and recommended daily fiber intake of >25g.
‡Based on 8.4 MJ/day (2,000 kcal/day) diet and recommended daily sodium intake of <2g.

and other food products sold in UK. In addition to the the nutrients per portion or per 100g, a traffic light system (red, amber, green) is used to inform the consumer, whether the meal is healthy (green) or unhealthy (red) with respect to a given property. We employed these guidelines in Table 2. As the FSA scoring system is rather unpractical to use in a recommender scenario, one might want to use a single metric by following the procedure proposed by Sacks, Rayner and Swinburn (2009) who first assign an integer value to each color (green=1, amber=2 and red=3) then sum the scores for each macro nutrient, resulting in a final range from 4 (very healthy) to 12 (very unhealthy). A further health index, which may offer utility is the 'Healthy Eating Index' (HEI, 2016) proposed by the USDA. The index was developed to target the US population. To date it has not been applied in any food recommender systems project.

Other useful resources for building food recommender systems are provided by FOODSUBS[17], a food thesaurus service which can suggest food substitutes. This might be helpful to implement food recommender systems promoting healthier eating (see (Achananuparp & Weber, 2016)) by replacing unhealthy ingredients in a meal with more healthy variants, but also assist people with allergies or intolerances.

Food word lists, such as provided by ENCHANTEDLEARNING[18] and WIKIPEDIA[19] provides a rich knowledge base relating to food and cooking and may be used to assist with the normalization process of ingredients.

Finally, one may also employ health data as provided by the Centers for Disease Control and Prevention (CDC) in the US. The reports contain state and county data of diabetes and obesity. As different regions have different impact on what and how people eat (see (Trattner, Parra & Elsweiler, 2017)), this might be a useful source of information when implementing food recommender systems for different regions and areas in the US (Said & Bellogín, 2014).

Food Recommender System Frameworks

To date, research in the food recommender systems domain relies mostly relies on software custom built by researchers themselves explicitly for the purpose of their research. To the best of our knowledge, there is no food recommender systems framework available that has been shared by the research community or on open-access platforms, such as GITHUB[20]. This makes it challenging not only to progress the research in that area, but also to reproduce or validate findings published already. To counter this trend, in our own research, we have recently started to use publicly available frameworks, such as the well-known LibRec library. The framework is implemented in the Java programming language and comprises a relatively complete set of standard recommender systems algorithms, such as UserKNN, ItemKNN, BPR, SVD++, and so on, to tackle the rating prediction and item ranking problem. In (Trattner & Elsweiler,

[17] http://www.foodsubs.com

[18] http://www.enchantedlearning.com/wordlist/food.shtml

[19] http://www.wikipedia.org

[20] http://www.github.com

2017) we adopted the framework with pre- and post-filtering functions (as described in the previous Section) to re-rank items (in our case) recipes in terms of their healthiness. We are happy to share this code upon request. The framework can also be easily extended to the problem of recommending, e.g., recipes to a group of people as well as generating personalized meal plans. Other examples of frameworks in other programming languages may be found on Graham Jenson's Github page[21] as well as on the RecSys Wiki[22].

Historical Evolution and Versions of the System

The earliest examples of food recommender systems were proposed by the case-based reasoning (CBR) community (Hammond, 1986; Hinrichs, 1989). In contrast to current state-of-the-art food recommender approaches both employed planning algorithms taking a set of queries e.g. groceries as input to generate meal plans or a single new recipe. Technically speaking these systems bear little relation to modern systems. Later, systems emerged employing simple variants of today's well-known content-based and collaborative filtering recommender algorithms. Examples include, for instance the works of Aberg (2006); Lawrence et al. (2001); Mankoff et al. (2002).

The first food recommenders built which are directly comparable to modern systems, i.e. which employ standard algorithms such as UserKNN was presented in Berkovsky and Freyne (2010); Freyne and Berkovsky (2010). These were the first examples, where recipe datasets were used as a basis and the system was reliably evaluated. Subsequently other works emerged employing more advanced techniques to recommend food to people. Examples include the work of (van Pinxteren et al., 2011), which was the first to derive a similarity metric for recipes to be used for recommending healthful meals; (Ueta et al., 2011) and (El-Dosuky et al., 2012), which employ knowledge-base food recommendation approaches; and (Kuo et al., 2012) which employs tags to derive a knowledge graph to connect recipes and exploit this graph for recommending menus.

Other break through work was performed by (Teng et al., 2012), who proposed the use of ingredient networks to produce recommendations or the work of Harvey et al. (2013), who proposed a model accounting for food selection biases.

A significant break-through was recently made by Yang et al. (2017) who were able to develop a constraint-based (with different types of diets) mobile food recommender system exploring food images to learn about user food preferences. All previous approaches had relied on ratings or to some extent on tags (Ge, Elahi et al., 2015).

Behavior-based investigations, which go beyond the classic food recommender systems papers can also be considered to have progressed the field. We include our own work showing that people typically prefer the unhealthy recipes in this bracket (Trattner & Elsweiler, 2017). This was the first study in the context that deals with the health-aware recipe recommender systems problem. Other work in this direction include (Trattner, Rokicki & Herder, 2017) (not shown in Table 1) and (Rokicki et al., 2016) which illustrate differences in online food consumption with respect to hobbies and gender.

Finally, we would like to highlight our most recent work (Elsweiler, Trattner & Harvey, 2017) which investigated to which extent food recommender can nudge people towards healthier food choices.

Evaluation: Metrics and Methodologies

The methods of evaluation applied to food recommender systems have evolved over time. The early concept papers found in the literature do not employ any kind of evaluation (Hammond, 1986; Hinrichs & Kolodner, 1991). With the work of (Freyne & Berkovsky, 2010) researchers started to employ evaluation techniques recognized by the community today as standard practices (Herlocker, Konstan, Terveen & Riedl, 2004a; Ricci et al., 2011).

The most commonly taken approach (as can be seen in the summarized literature in Table 1) is to perform simulations using historical data (see Section 'Implementation Resources'). The experimental design specifics vary, but typically datasets are split into training and testing subsets to mimic user-profiles and feedback given for recommendations. Similar to other recommender domains, historical datasets are typically split such that 80% of the data is used for training with the remaining 20% held-out for testing. Alternatives are to use k-fold validation (Harvey et al., 2013; Trattner & Elsweiler, 2017) or leave-one out protocol (Freyne & Berkovsky, 2010). The exact means by which collections are sourced varies from using naturalistic collections crawled from the web (Trattner & Elsweiler, 2017) or from donated sources (Trevisiol et al., 2014) to running user studies to collect small sets of data (Harvey et al., 2013).

Different metrics have been applied to measure the performance of algorithms in such systems. These typically reflect the error in the predicted ratings (Freyne & Berkovsky, 2010; Harvey et al., 2013) e.g. Mean Absolute Error (MAE) or Root Spare Mean Error (RSME) or the quality of the top-n ranked list of items e.g. Recall, Precision, Mean Average Precision (MAP) and Normalized Discounted Cumulative Gain (NDCG) (Herlocker, Konstan, Terveen & Riedl, 2004b).

Mirroring the developments in the recommender systems community generally, earlier contributions focused on the rating prediction task whereas more recent and current work treats recommendation as a ranking problem (e.g., (Cheng et al., 2017; Trattner & Elsweiler, 2017; Yang et al., 2017)).

[21] https://github.com/grahamjenson/list_of_recommender_systems

[22] http://www.recsyswiki.com/wiki/Recommendation_Software

Assessing the accuracy of recommendation is typically not enough for recommender systems and in food recommenders is no exception. Diversity of ingredients used in profiles was measured using Simpson and simple diversity metrics (Elsweiler & Harvey, 2015).

Incorporating health-aspects in the process requires additional metrics to be defined. As our own work shows, see (Trattner & Elsweiler, 2017), metrics derived from the guidelines published by governmental bodies or health organizations are appropriate.

In addition to calculating a mean over all food items recommended on per user basis (see Table 2), we additionally introduced two further measures referred to as ΔFSA and ΔWHO, which capture the difference in healthfulness between test set items of a user and actual predicted items, as shown in the formulae below

$$\Delta WHO = \sum_{u=1}^{|U|} \left(\sum_{i=1}^{|Train_u|} who_i - \sum_{j=1}^{|Pred_u|} who_j \right) \tag{4}$$

$$\Delta FSA = \sum_{u=1}^{|U|} \left(\sum_{i=1}^{|Train_u|} fsa_i - \sum_{j=1}^{|Pred_u|} fsa_j \right) \tag{5}$$

, where $|U|$ denotes the total number of users in the dataset, $|Train_u|$ the size of the train set for user u respectively, $|Pred_u|$ the size of the set for the predicted items and who_i, who_j and fsa_i, fsa_j represent the WHO, FSA health scores for items (i and j) in these sets.

These delta measures are useful as they capture whether the recommended items are more or less healthy than those already rated positively by the user. The same procedure can also be applied to calculate a delta between the test and prediction sets to observe whether the recommended items are actually more or less healthy to what the user would actually eat in the future.

Similar to other recommender domains, studies employing online evaluation protocols, such as A/B testing or laboratory studies for the purpose of testing the performance of food recommender systems are rare. Among the studies to employ online testing is for instance the work of (Freyne, Berkovsky, Baghaei et al., 2011) who ran two types of meal planners in a live system. The two methods tested were a personalized and a non-personalized algorithm. Over the course of 12 weeks over 5000 users participated in the study. According to the authors an A/B like setup was chosen to refer half of the users to the personalized condition and half of the user to the non-personalized one. Earlier work from the same authors employed also some variant of online experiment to gather ratings from users on recipes by e.g. using Amazon's Mechanical Turk platform (Freyne & Berkovsky, 2010; Freyne, Berkovsky & Smith, 2011). However, rather than generating the recommendations on the fly to test their validity, in the end, an offline protocol was utilized. A recent study by Yang et al. (2017) employed not only offline testing but also an online study protocol to evaluate a mobile food recommender system. In particular they recruited 60 participants through the university mailing list, Facebook, and Twitter. The study, conducted as a online Web service, consisted of three phases. First, each participant was questioned on any dietary restrictions that may apply, such as the need to avoid gluten. Second, each user was asked to express their preferences by highlighting images of food they find appealing. Lastly, 20 meal recommendations were generated of which 10 were shown in a random order and 10 as proposed by the the authors' "Yum-me" algorithm. The participants had the task of classifying the 20 recipes as to whether it is appealing or not.

A final work worthy of mention is an online study that has been recently conducted by the authors with the goal of investigating the potential to nudge people towards healthier food choices via recommendations (Elsweiler, Trattner & Harvey, 2017). The work employed three online studies. Similar to the previously mentioned work we implemented a Web service and recruited between 107 and 138 participants per study. By varying the amount of information shown about two algorithmically determined similar recipes, we were able to learn about the choices people make, the users' perception of these recipes and what influenced these. By applying machine learning approaches we were able to predict with relative certainty, which recipe of the two participants would prefer and demonstrate that the models developed can be used to influence the choices made.

In summary, no specialized offline protocols exist for the evaluation food recommender systems. Typically standard metrics are used to determine prediction accuracy and diversity. Furthermore, no standardized or specialized online evaluation protocols exist for food recommender systems. Current approaches rely on methods that have been previously developed in other recommender domains such as movies or music. Exceptions are the metrics specifically designed to incorporate healthy nutrition into the process, such as the WHO and FSA scores in (Trattner & Elsweiler, 2017).

Lessons Learned and Future Directions

Thus, food recommendation is an important domain both for individuals and society. What the work described in this paper shows is that despite its importance, food item recommendation, in comparison to other domains is relatively under-researched. The work that has been performed to date shows that although user taste predictions for food can be achieved with existing methods, the performance achieved is poorer than in other domains.

This means that preference learning is should remain a focus for the food domain because experiments described in the literature have shown that even regardless of the source of user feedback applied (i.e. ratings, tags or comments) standard methods are only capable of producing relatively unsatisfactory performance. It is clear that new methods are required for the food domain and some work has shown promise. Yang and colleague's (2017) work uses images and embeddings (DNNs) to learn user preferences and the results are very promising.

Other key findings in the literature relating to preference prediction are those illustrating the importance of context variables. One promising research direction would be to capture important context variables via different sensors and incorporate these into recommendation models.

Relating to context, social situations and recommendation for groups needs to be considered more concretely. The pervasiveness of social culinary experiences and how these influence food choices need to be considered by technological systems.

One particular task in food recommender systems, which for societal and socio-economic reasons, has become a hot research focus is food recommenders for nutritional health. Researchers have proposed diverse methods of incorporating nutrition (nutritional components in algorithm, meal plans, and nudging), but to date all of these proposals remain preliminary and it is not yet clear, which is the best approach to take.

As a final note, one further aspect which needs to develop in the community is the evaluation of food recommenders and the methods employed to do so. In the literature evaluation has mostly been offline with proprietary collections. As a community we need to work together to achieve standard data collections, standard base-line approaches and importantly, more online studies to understand how our approaches work as live systems used in naturalistic scenarios.

References

Abbar, S., Mejova, Y. & Weber, I. (2015). You tweet what you eat: Studying food consumption through twitter. In *Proceedings of the 33rd annual acm conference on human factors in computing systems* (pp. 3197–3206).

Aberg, J. (2006). Dealing with malnutrition: A meal planning system for elderly. In *Aaai spring symposium: Argumentation for consumers of healthcare* (pp. 1–7).

Achananuparp, P. & Weber, I. (2016). Extracting food substitutes from food diary via distributional similarity. *CoRR*, *abs/1607.08807*. Retrieved from http://arxiv.org/abs/1607.08807

Beaglehole, R. (2016, October). *Misunderstaning vs reality.* Retrieved from http://www.who.int/chp/advocacy/MediaFeatures_EN_web.pdf

Berkovsky, S. & Freyne, J. (2010). Group-based recipe recommendations: analysis of data aggregation strategies. In *Proceedings of the fourth acm conference on recommender systems* (pp. 111–118).

Cheng, H., Rokicki, M. & Herder, E. (2017). 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). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/3079628.3079641 doi: 10.1145/3079628.3079641

Chokr, M. & Elbassuoni, S. (2017). Calories prediction from food images. In *Aaai* (pp. 4664–4669).

De Choudhury, M., Sharma, S. & Kiciman, E. (2016). Characterizing dietary choices, nutrition, and language in food deserts via social media. In *Proceedings of the 19th acm conference on computer-supported co-operative work & social computing* (pp. 1157–1170). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/2818048.2819956 doi: 10.1145/2818048.2819956

Elahi, M., Ge, M., Ricci, F., Massimo, D. & Berkovsky, S. (2014). Interactive food recommendation for groups. In *Recsys posters.*

El-Dosuky, M., Rashad, M., Hamza, T. & El-Bassiouny, A. (2012). Food recommendation using ontology and heuristics. In *International conference on advanced machine learning technologies and applications* (pp. 423–429).

Elsweiler, D. & Harvey, M. (2015). Towards automatic meal plan recommendations for balanced nutrition. In *Proceedings of the 9th acm conference on recommender systems* (pp. 313–316). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/2792838.2799665 doi: 10.1145/2792838.2799665

Elsweiler, D., Harvey, M., Ludwig, B. & Said, A. (2015). Bringing the" healthy" into food recommenders. In *Dmrs* (pp. 33–36).

Elsweiler, D., Hors-Fraile, S., Ludwig, B., Said, A., Schäfer, H., Trattner, C., … Valdez, A. C. (2017). Second workshop on health recommender systems: (healthrecsys 2017). In *Proceedings of the eleventh ACM conference on recommender systems, recsys 2017, como, italy, august 27-31, 2017* (pp. 374–375). Retrieved from http://doi.acm.org/10.1145/3109859.3109955 doi: 10.1145/3109859.3109955

Elsweiler, D., Ludwig, B., Said, A., Schäfer, H. & Trattner, C. (2016). Engendering health with recommender systems. In *Proceedings of the 10th ACM conference on recommender systems, boston, ma, usa, september 15-19, 2016* (pp. 409–410). Retrieved from http://doi.acm.org/10.1145/2959100.2959203 doi: 10.1145/2959100.2959203

Elsweiler, D., Trattner, C. & Harvey, M. (2017). Exploiting food choice biases for healthier recipe recommendation. In *Proceedings of the 40th international acm* *sigir conference on research and development in information retrieval* (pp. 575–584). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/3077136.3080826 doi: 10.1145/3077136.3080826

Ferguson, E. L., Darmon, N., Briend, A. & Premachandra, I. M. (2004). Food-based dietary guidelines can be developed and tested using linear programming analysis. *The Journal of nutrition*, *134*(4), 951–957.

Freyne, J. & Berkovsky, S. (2010). Intelligent food planning: Personalized recipe recommendation. In *Proceedings of the 15th international conference on intelligent user interfaces* (pp. 321–324). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/1719970.1720021 doi: 10.1145/1719970.1720021

Freyne, J., Berkovsky, S., Baghaei, N., Kimani, S. & Smith, G. (2011). Personalized techniques for lifestyle change. *Artificial Intelligence in Medicine*, 139–148.

Freyne, J., Berkovsky, S. & Smith, G. (2011). Recipe recommendation: Accuracy and reasoning. In *International conference on user modeling, adaptation, and personalization* (pp. 99–110).

FSA. (2016). Guide to creating a front of pack (fop) nutrition label for pre-packed products sold through retail outlets. available at https://www.food.gov.uk/sites/default/files/multimedia/pdfs/pdf-ni/fop-guidance.pdf. last accessed on 20.6.2016.

Ge, M., Elahi, M., Fernaández-Tobías, I., Ricci, F. & Massimo, D. (2015). Using tags and latent factors in a food recommender system. In *Proceedings of the 5th international conference on digital health 2015* (pp. 105–112). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/2750511.2750528 doi: 10.1145/2750511.2750528

Ge, M., Ricci, F. & Massimo, D. (2015). Health-aware food recommender system. In *Proceedings of the 9th acm conference on recommender systems* (pp. 333–334). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/2792838.2796554 doi: 10.1145/2792838.2796554

Gibney, M., Vorster, H. & Kok, F. (2002). *Introduction to human nutrition*.

Griffiths, T. (2002). Gibbs sampling in the generative model of latent dirichlet allocation. , x–y.

Hammond, K. J. (1986). Chef: A model of case-based planning. In *Proceedings of aaai* (pp. 267–271).

Harvey, M. & Elsweiler, D. (2015). Automated recommendation of healthy, personalised meal plans. In *Proceedings of the 9th acm conference on recommender systems* (pp. 327–328).

Harvey, M., Ludwig, B. & Elsweiler, D. (2012). Learning user tastes: A first step to generating healthy meal plans. In *First international workshop on recommendation technologies for lifestyle change (lifestyle 2012)* (p. 18).

Harvey, M., Ludwig, B. & Elsweiler, D. (2013). You are what you eat: Learning user tastes for rating prediction. In *Proceedings of the 20th international symposium on string processing and information retrieval - volume 8214* (pp. 153–164). New York, NY, USA: Springer-Verlag New York, Inc. Retrieved from http://dx.doi.org/10.1007/978-3-319-02432-5_19 doi: 10.1007/978-3-319-02432-5_19

HEI. (2016, October). *Healthy eating index*. Retrieved from https://www.cnpp.usda.gov/healthyeatingindex

Herlocker, J. L., Konstan, J. A., Terveen, L. G. & Riedl, J. T. (2004a). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, *22*(1), 5–53.

Herlocker, J. L., Konstan, J. A., Terveen, L. G. & Riedl, J. T. (2004b, January). Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.*, *22*(1), 5–53. Retrieved from http://doi.acm.org/10.1145/963770.963772 doi: 10.1145/963770.963772

Hinrichs, T. R. (1989). Strategies for adaptation and recovery in a design problem solver. In *Proceedings of the 2nd workshop on case-based reasoning* (pp. 115–118).

Hinrichs, T. R. & Kolodner, J. L. (1991). The roles of adaptation in case-based design. In *Aaai* (Vol. 91, pp. 28–33).

Howard, S., Adams, J. & White, M. (2012). Nutritional content of supermarket ready meals and recipes by television chefs in the united kingdom: cross sectional study. *BMJ*, *345*, e7607.

Hu, Y., Koren, Y. & Volinsky, C. (2008). Collaborative filtering for implicit feedback datasets. In *Proc. of icdm'08* (pp. 263–272).

Khan, A. S. & Hoffmann, A. (2003). Building a case-based diet recommendation system without a knowledge engineer. *Artificial Intelligence in Medicine*, *27*(2), 155–179.

Kuo, F.-F., Li, C.-T., Shan, M.-K. & Lee, S.-Y. (2012). Intelligent menu planning: Recommending set of recipes by ingredients. In *Proceedings of the acm multimedia 2012 workshop on multimedia for cooking and eating activities* (pp. 1–6).

Kusmierczyk, T., Trattner, C. & Nørvåg, K. (2015). Temporal patterns in online food innovation. In *Proceedings of the 24th international conference on world wide web* (pp. 1345–1350).

Kusmierczyk, T., Trattner, C. & Nørvåg, K. (2016). Understanding and predicting online food recipe production patterns. In *Proceedings of the 27th acm conference on hypertext and social media* (pp. 243–248).

Lawrence, R. D., Almasi, G. S., Kotlyar, V., Viveros, M. & Duri, S. S. (2001). Personalization of supermarket product recommendations. In *Applications of data mining to electronic commerce* (pp. 11–32). Springer.

Mankoff, J., Hsieh, G., Hung, H. C., Lee, S. & Nitao, E. (2002). Using low-cost sensing to support nutritional awareness. In *International conference on ubiquitous computing* (pp. 371–378).

Mormann, M. M., Navalpakkam, V., Koch, C. & Rangel, A. (2012). Relative visual saliency differences induce sizable bias in consumer choice.

Müller, M., Mika, S., Harvey, M. & Elsweiler, D. (2012). Estimating nutrition values for internet recipes. In *Pervasive computing technologies for healthcare (pervasivehealth), 2012 6th international conference on* (pp. 191–192).

Ornish, D., Brown, S., Billings, J., Scherwitz, L., Armstrong, W., Ports, T., … Brand, R. (1990). Can lifestyle changes reverse coronary heart disease?: The lifestyle heart trial. *The Lancet*, *336*(8708), 129 – 133. Retrieved from http://www.sciencedirect.com/science/article/pii/014067369091656U

Park, M.-H., Park, H.-S. & Cho, S.-B. (2008). Restaurant recommendation for group of people in mobile environments using probabilistic multi-criteria decision making. In *Computer-human interaction* (pp. 114–122).

Ricci, F., Rokach, L. & Shapira, B. (2011). *Introduction to recommender systems handbook*. Springer.

Rokicki, M., Herder, E., Kuśmierczyk, T. & Trattner, C. (2016). Plate and prejudice: Gender differences in online cooking. In *Proceedings of the 2016 conference on user modeling adaptation and personalization* (pp. 207–215). New York, NY, USA: ACM. doi: 10.1145/2930238.2930248

Rokicki, M., Herder, E. & Trattner, C. (2017). How editorial, temporal and social biases affect online food popularity and appreciation. In *Icwsm* (pp. 192–200).

Sacks, G., Rayner, M. & Swinburn, B. (2009). Impact of front-of-pack âĂŸtraffic-lightâĂŹnutrition labelling on consumer food purchases in the uk. *Health promotion international*, *24*(4), 344–352.

Said, A. & Bellogín, A. (2014). You are what you eat! tracking health through recipe interactions. In *Rsweb@ recsys*.

Sano, N., Machino, N., Yada, K. & Suzuki, T. (2015). Recommendation system for grocery store considering data sparsity. *Procedia Computer Science*, *60*, 1406–1413.

Schäfer, H., Elahi, M., Elsweiler, D., Groh, G., Harvey, M., Ludwig, B., … Said, A. (2017). User nutrition modelling and recommendation: Balancing simplicity and complexity. In *Adjunct publication of the 25th conference on user modeling, adaptation and personalization* (pp. 93–96).

Schäfer, H., Hors-Fraile, S., Karumur, R. P., Calero Valdez, A., Said, A., Torkamaan, H., … Trattner, C. (2017). Towards health (aware) recommender systems. In *Proceedings of the 2017 international conference on digital health* (pp. 157–161). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/3079452.3079499 doi: 10.1145/3079452.3079499

Scheibehenne, B., Greifeneder, R. & Todd, P. M. (2010). Can there ever be too many options? a meta-analytic review of choice overload. *Journal of Consumer Research*, *37*(3), 409–425.

Schur, E., Kleinhans, N., Goldberg, J., Buchwald, D., Schwartz, M. & Maravilla, K. (2009). Activation in brain energy regulation and reward centers by food cues varies with choice of visual stimulus. *International journal of obesity (2005)*, *33*(6), 653.

Straßburg, A. (2010). Ernährungserhebungen - methoden und instrumente. *Ernährungs Umschau*.

Teng, C.-Y., Lin, Y.-R. & Adamic, L. A. (2012). Recipe recommendation using ingredient networks. In *Proceedings of the 4th annual acm web science conference* (pp. 298–307).

Trattner, C. & Elsweiler, D. (2017). 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).

Trattner, C., Elsweiler, D. & Howard, S. (2017). Estimating the healthiness of internet recipes: A cross sectional study. *Frontiers in Public Health*, *5*, 16. Retrieved from http://journal.frontiersin.org/article/10.3389/fpubh.2017.00016 doi: 10.3389/fpubh.2017.00016

Trattner, C., Parra, D. & Elsweiler, D. (2017). Monitoring obesity prevalence in the united states through bookmarking activities in online food portals. *PloS one*, *12*(6), e0179144.

Trattner, C., Rokicki, M. & Herder, E. (2017). 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). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/3099023.3099072 doi: 10.1145/3099023.3099072

Trevisiol, M., Chiarandini, L. & Baeza-Yates, R. (2014).

Buon appetito: Recommending personalized menus. In *Proceedings of the 25th acm conference on hypertext and social media* (pp. 327–329). New York, NY, USA: ACM. Retrieved from `http://doi.acm.org/10.1145/2631775.2631784` doi: 10.1145/2631775.2631784

Ueta, T., Iwakami, M. & Ito, T. (2011). A recipe recommendation system based on automatic nutrition information extraction. In *Proceedings of the 5th international conference on knowledge science, engineering and management* (pp. 79–90). Berlin, Heidelberg: Springer-Verlag. Retrieved from `http://dx.doi.org/10.1007/978-3-642-25975-3_8` doi: 10.1007/978-3-642-25975-3_8

van Pinxteren, Y., Geleijnse, G. & Kamsteeg, P. (2011). Deriving a recipe similarity measure for recommending healthful meals. In *Proceedings of the 16th international conference on intelligent user interfaces* (pp. 105–114). New York, NY, USA: ACM. Retrieved from `http://doi.acm.org/10.1145/1943403.1943422` doi: 10.1145/1943403.1943422

Wagner, C., Singer, P. & Strohmaier, M. (2014). The nature and evolution of online food preferences. *EPJ Data Science*, *3*(1), 1.

Wansink, B. (2006). *Mindless eating*. Bantam Books.

West, R., White, R. W. & Horvitz, E. (2013). From cookies to cooks: Insights on dietary patterns via analysis of web usage logs. In *Proceedings of the 22nd international conference on world wide web* (pp. 1399–1410).

WHO. (2003). Diet, nutrition and the prevention of chronic diseases. *World Health Organ Tech Rep Ser*, *916*(i-viii).

Yang, L., Cui, Y., Zhang, F., Pollak, J. P., Belongie, S. & Estrin, D. (2015). Plateclick: Bootstrapping food preferences through an adaptive visual interface. In *Proceedings of the 24th acm international on conference on information and knowledge management* (pp. 183–192).

Yang, L., Hsieh, C.-K., Yang, H., Pollak, J. P., Dell, N., Belongie, S., … Estrin, D. (2017). Yum-me: A personalized nutrient-based meal recommender system. *ACM Transactions on Information Systems (TOIS)*, *36*(1), 7.

Zhang, Z., Yang, Y., Yue, Y., Fernstrom, J., Jia, W. & Sun, M. (2011). Food volume estimation from a single image using virtual reality technology. In *Bioengineering conference (nebec), 2011 ieee 37th annual northeast.*

Zhu, Y.-X., Huang, J., Zhang, Z.-K., Zhang, Q.-M., Zhou, T. & Ahn, Y.-Y. (2013). Geography and similarity of regional cuisines in china. *PloS one*, *8*(11), e79161.