Deriving a recipe similarity measure for recommending healthful meals.

CONFERENCE PAPER · JANUARY 2011

DOI: 10.1145/1943403.1943422 · Source: DBLP

3 AUTHORS, INCLUDING:

Gijs Geleijnse
Philips
35 PUBLICATIONS **194** CITATIONS
SEE PROFILE

Paul Kamsteeg
Radboud University Nijmegen
5 PUBLICATIONS **27** CITATIONS
SEE PROFILE

Deriving a Recipe Similarity Measure for Recommending Healthful Meals

Youri van Pinxteren[1,2], Gijs Geleijnse[1] and Paul Kamsteeg[2]

[1] Philips Research Europe
High Tech Campus 34
Eindhoven, the Netherlands
gijs.geleijnse@philips.com

[2] Radboud University
Department of Artificial Intelligence
Nijmegen, the Netherlands
p.kamsteeg@donders.ru.nl

ABSTRACT

A recipe recommender system may stimulate healthful and varied eating, when the presented recipes fit the lifestyle of the user. As consumers face the barrier to change their eating and cooking behavior, we aim for a strategy to provide more healthful variations to routine recipes. In this paper, a similarity measure for recipes is derived by taking a user-centered approach. Such a measure can be used to recommend healthier alternatives to commonly selected meals, which are perceived to be similar. Recipes presented using this strategy may fit the demand for health and variation within the boundaries of a busy lifestyle. Having derived and evaluated a recipe similarity measure, we explore its use through an at-home trial.

Author Keywords

Recipe similarity, card-sorting, natural language processing, user evaluation, at-home trail.

ACM Classification Keywords

H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION

Overweight and obesity are internationally recognized problems that cause many chronic diseases. Proper food is an important factor to improve such unhealthy conditions [30]. Although consumers are in general aware of improper eating habits, they often do not comply with corresponding behaviors. Their food choice is influenced by many factors [30], but unwillingness to spend cognitive effort on food preparation as well as a lack of time are important barriers to achieving a healthy diet [7]. Hence, effective interventions to promote healthful home-cooking should address these barriers.

Apart from selecting meals that do not meet the recommended dietary, the "nutritional gatekeeper" of the family [26] generally also selects meals and ingredients with limited variation. This is especially the case amongst younger adults [9] and lower income families [21]. In [22], People have the desire to bring more variety into their eating pattern [15], but studies also showed that people in general only try out recipes on special occasions: only 22% report to do so more frequently than twice a month [22]. Moreover, Twigt's study demonstrates that it is commonly recognized that variation is part of a healthful diet. Hence, there is both a need for support to select healthful meals as well as to expand the repertoire of recipes. Thus, there is a need for a solution that provides people with a larger repertoire of healthful meals that fit within their lifestyle.

Studies indicate that informing people of the health benefits of a meal is not enough to promote the selection of healthier alternatives [28]. Through a field study, Wansink and Chandon [27] found that the use of *low fat* labels actually leads to an increase of calorie intake, as people increased their serving sizes and consumptions. Downs et al. [8] found that caloric information has only a small positive effect on food selections in restaurants. Moreover, the study also showed `perverse' negative effects, like promoting calorie-rich meals among dieters. As perceived self-control is an important prerequisite for effective interventions, Wisdom et al. [28] conclude that the use of subtle influencing strategies, or nudges [20], may be more effective than the presentation of information alone.

A generally recognized approach to promote health-aware behavior is computer tailoring, where the information and advice are personalized to the user [6]. Research shows that computer tailored advice is more effective to educate people about nutrition than general information.

The recommendation of healthful meals that are perceived as attractive can be seen as a nudge to make healthful alternatives more salient. By offering attractive recipes rather than general nutritional information, actionable advice [2] is given. If this advice is tailored to personal preferences such that the meal fits the daily routine, the perceived barriers for preparing healthy foods are targeted. By providing a wide variety of recipes and allowing the user to also select less healthy alternatives, compliance to the system is promoted [3].

In this paper, we present the design of a similarity measure for recipes. Such a measure can be used to recommend

IUI 2011, February , 2011, Palo Alto, CA, USA.

healthier alternatives to commonly selected meals, which are *perceived* to be similar. Such alternatives may for example contain more vegetables or less saturated fat. Using a recipe similarity measure, recipes that are small variations to familiar ones can be identified. Such a strategy would fit the demand for health and variation within the boundaries of a busy lifestyle. Moreover, a recipe similarity measure can be used to identify differences between meals on consecutive days, enhancing a varied diet over a longer period of time.

In constructing the recipe similarity measure presented in this paper, a user-centered approach was taken. As attitudes and beliefs about food and recipes are culturally and regionally dependent [28], these considerations should also be taken into account when designing a model for a content-based recipe recommender. Although the model presented in this work may specifically apply to the Dutch situation, the method of deriving and applying the measure can be reused for other audiences and languages.

In a small scale at-home study, the use of the recipe similarity measure is explored. This study provides valuable first insights in the practical use of a similarity-based recipe recommender system.

RELATED WORK

Recipe recommender systems have received early attention in the field of case-based reasoning. In the CHEF system [14], a recipe is modeled based on its ingredients, taste (e.g. hot or fresh), texture and type of dish. However, his approach requires an extensive knowledge base and it is unclear whether elements like taste and texture can be automatically derived from a recipe text.

A recipe navigation system with social recommendations is studied by Svensson et al. [19]. A collaborative filtering approach is applied, based on rating for shared recipes. The system was evaluated in a large scale user trial. Although the performance of the recommender system was not formally evaluated, usage and post-trial interviews revealed that this functionality was well-appreciated. However, where performance is expected to increase towards the end of the trial, the usage did not increase.

Content-based recipe recommender systems have been recently addressed by Freyne and Berkovsky [11]. The recommendations of recipes in this approach are based on user ratings for recipes. Strategies are presented to mitigate ratings for recipes to the included ingredients. Based on the predicted ratings for individual ingredients, ratings for other recipes are subsequently computed. Their results show that this approach outperforms a direct collaborative filtering approach on the given food items.

Zhang et al. [32] also make use of an ingredient representation. But, where Freyne and Berkovsky make the assumption that the perceived similarity between recipes directly corresponds to the ingredients present in a recipe. All ingredients are assumed to be of equal importance, Zhang et al. distinguish three levels of importance for the ingredients, which are manually assigned. Using this mechanism, ingredients that are considered by the researchers to be more important, have the largest contribution to the similarity score. In [32], a hierarchical representation (using WordNet and other structured sources) is used to match ingredients. Using this mechanism, the similarity between for example two different kinds of pasta can be detected.

Wang et al. [25] created a graph-based model of recipes using ingredients and cooking directions. First, they semi-automatically represent the recipes as graphs with objects (i.e. the ingredients) and actions (e.g. stirring, frying). To express the similarity between two recipes, the corresponding graphs are compared. This measure was used to distinguish Chinese regional dishes (Guangdong style vs. Sichuan style), which are known to have a different cooking procedure.

In our work, we revisit the process of recipe modeling by taking a user-centered approach, which, to the best of our knowledge, has never been done yet. We hence do not make any a priori assumption about the characteristics that determine the perceived similarity, such as ingredients or directions. Moreover, we do not assume any rating to be assigned by the users, avoiding the infamous cold-start problem and the need of an active user community.

IDENTIFYING SIMILARITY CHARACTERISTICS USING CARD-SORTING

Earlier work on recipe similarity measures focuses on models of the recipes using their ingredients [11] or ingredients and preparation steps [25]. Our goal is to derive a model for recipes taking a user-centered approach. To this end, a card-sorting experiment was designed to reveal the most important aspects that contribute to the perceived similarity between recipes for a Dutch audience.

Card sorting has been shown to be an effective technique to elicit groupings of concepts from untrained participants and non-experts [10]. In for example [16], card-sorting been successfully applied to food-related domains, but to our best knowledge, never to recipes for main courses.

Recipe Sets

Two sets of recipe cards were developed for this experiment. The cards contained the essential information about the recipe, such as the title, ingredients, cuisine, preparation time and preparation technique as well as a picture (viz. Figure 1). All recipes were popular recipes for main courses selected from a large Dutch recipe website. Two sets of recipe cards were created. The first set ("diverse set") of 66 cards was selected in such a way that the characteristics presented on the cards varied greatly. The recipes varied among meal type (e.g. salad, hotchpotch and stir-fry), starch-rich ingredients (rice, pasta etc.), meat, fish or vegetarian, preparation time and cuisine. The second set ("pasta set") consisted of twenty Italian pasta dishes with meat. This set was designed to obtain insights into the

factors that influence the more fine-grained differences between similar recipes.

Figure 1 Example of a recipe card as used in the card sorting task.

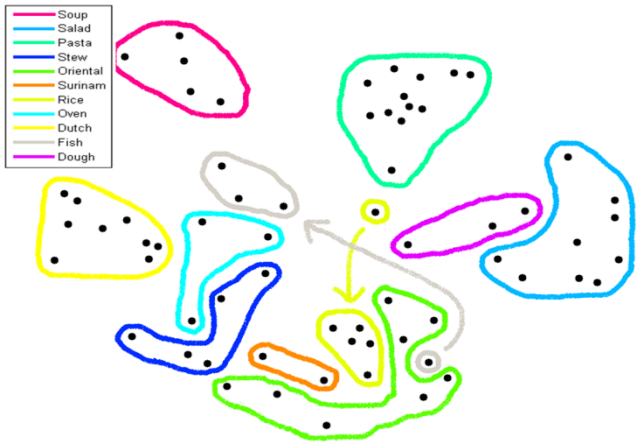

Figure 2. The MDS visualization of the distances between the recuoes in the *diverse* set.

Procedure

Fourteen Dutch participants (one male) conducted the card-sorting tasks. All participants indicated to cook regularly. In two free-sorting [10] tasks, the participants clustered the cards from both sets. After having formed groups of cards, in the verbalization step, the participants were invited to characterize the groups formed.

In addition to free-sorting, the participants were given a conditional ranking task, where one recipe in the Italian pasta set was selected as an anchor recipe. The participants were invited to rank the remaining recipes on a 5 point Likert scale (from very similar to very dissimilar).

Analysis and Results

Participants grouped the diverse set of 66 recipes into an average of 9.4 groups. The minimal number of groups used was six, while the maximum was thirteen. The distinct descriptions given in the verbalization step, 131 in total, were analyzed and (near) synonyms were merged.

The groupings of the recipes were visualized using Multi Dimensional Scaling (MDS, Figure 2) and Hierarchical Clustering Analysis (HCA, using Burton's F). Both techniques showed similar clusters, corresponding to quite diverse user descriptions like 'soup', 'oriental, 'pasta', 'stew' and 'fish'. The only recipes that were differently grouped by the two visualization techniques were Catalan paella (a dish with both fish and rice) and Chicken in Italian pasta sauce (both chicken and pasta).

Analysis shows that the perceived similarity between recipes is determined by a combination of higher level characteristics. The most prominent characteristic was *meal type*, where the corresponding six values were mentioned a total of 37 times. It turned out that participants used eight different characteristics to sort the diverse recipe set (viz. Table 1).

Figure 3. The MDS visualization of the distances between the Italian pasta recipes, using multidimensional scaling with Burton's F.

Table 1. All characteristics for diverse recipe set.

characteristic	Freq.	# values	Example values
Meal type	37	6	Salad, soup
Cuisine	32	16	Italian, Chinese
Starch	29	5	pasta, rice
Meat/fish/vegetarian	26	8	fish, turkey
technique	15	2	oven, stew
preparation time	7	2	short, long
heaviness	3	2	light, strong
healthfulness	1	1	unhealthy

Table 2. The characteristics for the Italian pasta set.

characteristic	Freq	# values	Example values
Meat	37	10	Salami, lamb
Starch	32	10	Spaghetti, penne
technique	29	1	oven
meal type	26	1	salad
vegetables	15	2	many, few
sauce	7	2	meat/tomato sauce
heaviness	3	2	light, strong
preparation time	1	1	long, short
ingredients	2	2	many, few

In correspondence to [11], [23] and [29], most characteristics are related to ingredients and preparation techniques. However, cuisine, meal type and preparation time are also of importance. Only three ('all kinds', 'special dishes' and 'other') out of 131 descriptions could not be converted into values of any of the eight characteristics.

The visualizations show that the participants considered all pasta recipes in the diverse set to be alike (which some participants also explicitly brought up during the Free Sorting task). This indicates that this recipe set only reveals the most influential characteristics of a recipe. To obtain a more fine-grained recipe model of the recipes, we analyzed the outcome of the sorting tasks for the Italian pasta set. In this set, some important characteristics are constrained: starch and cuisine are limited to pasta and Italian respectively, while meat/fish/vegetarian is limited to meat.

Participants clustered the twenty Italian pasta recipe cards into an average of 5.7 recipes per group. The minimal number of groups used was two (oven dish or not), while the maximum was eleven. Using HCA and MDS, visualizations of distances between the recipes were generated for the pasta set as well (viz. Figure 3), using the users' descriptions for clustering.

Again, all participant descriptions were analyzed and converted into recipe characteristics and their values (viz. Table 2). Now, eight (out of 80) descriptions could not be converted. This may indicate that participants had more difficulties sorting this set. Most of the unconverted descriptions received either uninformative labels, like 'pasta with meat' or simply incorrect ones, like 'vegetarian pasta' (note that all recipes in the set are pastas with meat). It turned out that participants used nine different characteristics to sort the recipe set.

The characteristics '*kind of meat'* (e.g. chicken, pork, ham, etc.) and '*kind of pasta*' (e.g. spaghetti, lasagna, stuffed pasta, etc.) have by far the highest occurrence, but participants mentioned a lot of different possible values. Hierarchical relations can be used for these characteristics to make them stronger. For example, participants mentioned values like ham, salami, sausage and minced meat, but also the more general values pork, chicken and beef.

Table 3. The 55 features divided over 13 characteristics.

Meal type: Soup, Salad, Hotchpotch
Cuisine: Dutch, French, Mediterranean, Italian, Spanish, Greek, Oriental, Chinese/Indonesian, Indian, Japanese, Thai, Mexican
Preparation time: Long, Short
Preparation technique: Oven, Stew
Starch: Potato, Pasta, Rice, Dough, Noodles, Bread
Kind of potatoes: Boiled, Baked, Fried, Mashed
Kind of pasta: Lasagna, Stuffed pasta, Short pasta, Long pasta
Kind of rice: Dry-boiled/Strong, Soft-boiled/Soft-sticky
Kind of dough: Flat bread, Pastry
Meat, Fish or seafood, Vegetarian
Kind of meat Poultry, Pork, Ham, Beef, Lamb, Mixed or processed
Kind of fish or Seafood White fish, Red fish, Fried or processed fish, Seafood
Vegetarian ingredients Meat substitute or pulse crop, Egg, Cheese, Dairy
Vegetables Many vegetables

FEATURE REPRESENTATION AND EXTRACTION

The card sorting task with the diverse set of recipes revealed the main characteristics that are used to assess similarity between recipes. The second part of the experiment, using the pasta set, only revealed criteria that are used to cluster this particular category of recipes. In order to apply to other recipe categories as well, we manually extended the set of values by including items that are analogous to the values and characteristics found with the Italian pasta set. The 55 selected features can be found in Table 3.

We hence have created a recipe representation that is quite different from other approaches in the literature. Instead of representing the recipes using their complete list of ingredients, only those ingredients are selected that are considered to be most important. Similarly, only a few preparation techniques are distinguishing. Our analysis also showed that other aspects of a recipe, such as the cuisine and the meal type, are important for this audience.

Having selected the 55 features, the goal is now to automatically convert a recipe into a (binary) feature-vector representation. As the features contain many generalized terms (e.g. red fish, dairy) that may not occur in the recipe text as such, a hierarchical representation of the ingredients is needed [32]. To this end we reused an ingredient hierarchy applied in earlier work [18].

An algorithm was created to determine the applicability for each of the features. For the ingredient-based features, it matches the ingredient list with the ingredient hierarchy. The cuisine and cooking technique are identified in the full recipe text using a small vocabulary of relevant terms (e.g. common synonyms for *hotchpotch*).

According to the participants in the card sorting study, a short preparation time means less than 30 minutes, while a long preparation time means at least one hour. These criteria are used to assess the short and long preparation time features. If a recipe does not mention an explicit preparation time, the sum of all mentioned partial preparation times is used.

Evaluating the feature extraction algorithm

We applied the feature extraction algorithm to a large set of 6886 recipe texts from a popular Dutch website. To assess the quality of the algorithm for this recipe set, we compared a sample of our algorithm results with the labeling of three volunteers.

To this end, forty recipes were randomly selected. The participants were asked to fill in the presence or absence of all 55 features in the forty recipes, leading to 2200 assignments per participant.

Table 4 shows performance of the algorithm compared to the assignments of the three participants. The table shows convincing results with an accuracy of around 94% (i.e. the true positives and true negatives). Analysis of the results showed that many mismatches could be explained by errors of the human assessors. For example, beans were not classified as pulse crops and the preparation technique stew was not recognized, even within a casserole recipe. Other differences were a result of differences in interpretation of the features. For example, some people find a salmon filet a processed fish. The property *many vegetables* and the values for cuisine also caused some differences. Finally, errors were made by the feature extraction algorithm. For example, in one recipe, salmon was not included in the ingredient list and not recognized.

Table 4. Automatic feature extraction for forty recipes compared to the ratings of three human judges.

	True positives	True negatives	False positives	False negatives
Participant 1	11.6%	83.8%	2.6%	2.0%
Participant 2	10.5%	82.8%	3.8%	3.0%
Participant 3	10.8%	83.8%	3.5%	2.0%
Average	11.0%	83.5%	3.3%	2.3%

We can conclude that our extraction method gives good results, as on average only 5.6% of the assignments differed in comparison with the assignments of three volunteers. Only a part of the errors are caused by actual assignment errors of the algorithm.

THE SIMILARITY MEASURE

Having defined a feature vector to represent a recipe and an algorithm to extract the features from the text, we now formulate a recipe similarity measure based on this feature vector. As the card sorting study showed that not every recipe property has the same influence on the perceived similarity between recipes, we assign a weight to each feature.

As the weights cannot be learned using the data currently at hand, we opted for a semi-automatic approach based on the user-generated data from the card-sorting tasks.

We used the basic assumption that a value (e.g. *salad*) is more important in proportion to the frequency with which it is mentioned. Further, we hypothesized that the frequency of the characteristic that the value corresponds to (e.g. *meal type*) is also of influence. This frequency would be the summed frequency of all corresponding values. Finally, we reasoned that the number of values that were actually mentioned for a characteristic is of importance, as this set of values is in fact a selection from a much larger set of possible values. Participants potentially could have mentioned many other possible values for the characteristics (e.g. baking was never mentioned as a preparation technique). This would mean that the less possible values mentioned, the more important those mentioned values are. Based on this analysis, we defined the following definition for the weight of value i in the feature vector.

$$w_i = \frac{c_i \sum_{j=0}^{n} c_j}{n}$$

where c_i is the number of times value i is mentioned in the card sorting study, and n is the number of different values that were mentioned for the characteristic to which value i belongs. For the added features without frequency data, the weight of the analogous "seed feature" was assigned.

Having assigned a weight to each of the features, we can compute the distance between two recipes x and y as the weighted Euclidean distance:

$$\mathrm{Dist}(x, y) = \sum_i w_i (x_i - y_i)^2$$

where w_i is the weight of feature i and x_i and y_i are Boolean values for recipes x and y.

EVALUATING THE SIMILARITY MEASURE

In order to assess the derived Recipe Similarity Measure (RSM), we want to compare it to the current state of the art in an evaluation with potential users. The goal is to identify an ordered list of recipes similar to an anchor recipe out of a large collection of recipes. To this end, recipe similarity measures as proposed in literature were studied. However, none of the encountered measures could be applied to our specific task. Where Wang et al. [25] used a semi-automatically derived representation of the recipes, we are interested in a fully automatic variant. Work in [11] assumes the use of user profiles with ratings for recipes, but such ratings are not available for any large collection of Dutch recipes. The similarity measure proposed by Zhang et al. [32] is a potential candidate. However, no evaluation

of their approach is known and the algorithm requires specific settings that are not discussed in the paper. Therefore, as no dedicated recipe similarity measure could be identified for benchmarking, we chose to compare the RSM to the Cosine Similarity Measure (CSM) [25], which is commonly used to determine the similarity between two texts.

In [11], Freyne and Berkovsky based their recipe recommender algorithms on the assessed level of attractiveness for a set of recipes. To investigate if the attractiveness of recipes also influences the perceived similarity between two recipes, we are also interested in the user rating for the recipes.

To this end, we conducted an online experiment, where participants were invited to rate the similarity between recipes as well as their attractiveness. We expect that the recipes identified by RSM are perceived to be more similar to an anchor recipe compared with the recipes identified by CSM. Moreover, it is expected that if an anchor recipe is rated as attractive, the most similar recipe in the collection also is.

Participants

Over a period of two weeks, 137 Dutch speaking participants (76 female) completed the web experiment. They were recruited through email and message boards. A kitchen appliance was raffled among the participants.

Procedure

After answering some basic demographic questions, the participants were presented with a series of twenty consecutive recipe pairs. For every pair, the participant was asked to rate the similarity on a 7-point Likert scale. He or she also had to indicate for every recipe how attractive they thought it was. For every recipe, we showed its title, cuisine, preparation time, ingredients and directions, all taken from the recipe text. An example screen shot of the web experiment can be seen in Figure X.

Design

We constructed four recipe sets, each one containing ten different anchor recipes. In this way, we were able to test 40 different recipes. These recipes were all randomly picked from the collection of over 6000 recipes for which we had extracted all recipe features.

Each 'anchor' recipe was both paired with the most similar recipe text found with the RSM and with the CSM. Using RSM, multiple recipes may receive the same highest similarity score. In this case, the CSM was used to force a decision among the recipe with the highest RSM score. This gave us twenty recipe pairs in each set.

The two methods were compared in a within-subject design. Participants were randomly assigned to one of the four recipe sets, and every participant was presented with all twenty pairs in that set (ten different anchor recipes, each occurring in two pairs with different "most similar" recipes). The participants were asked to rate the similarity of the recipes within each pair and in addition, to rate the attractiveness of every recipe. This implies that they were invited to rate twenty similarity scores and thirty attractiveness scores.

The order of anchor recipes was randomized for every participant. The two pairs with the same anchor recipe were presented consecutively, but the order of the two "most similar" recipes (RSM vs. CSM) was again randomized.

Results

First, to test our main hypothesis, whether the recipes found with RSM are perceived to be more similar than the recipes found with the baseline method, we conducted a matched pairs t-test. The results indicated that on average, participants rated the recipe identified with RSM more similar to the anchor recipe ($M = 4.08$, $SD = 0.027$) than the recipe found with CSM ($M = 3.37$, $SD = 0.027$), $t(136) = -12.937$, $p < .001$. The effect-size r of .74 indicates a large effect. This test confirms our hypothesis that the RSM strongly outperforms the CSM for this audience and the given set of recipes.

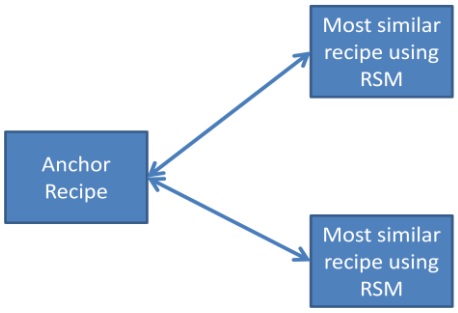

Figure 4. Set-up of web experiment: 10 anchor recipe are compared with the winning recipes using both methods.

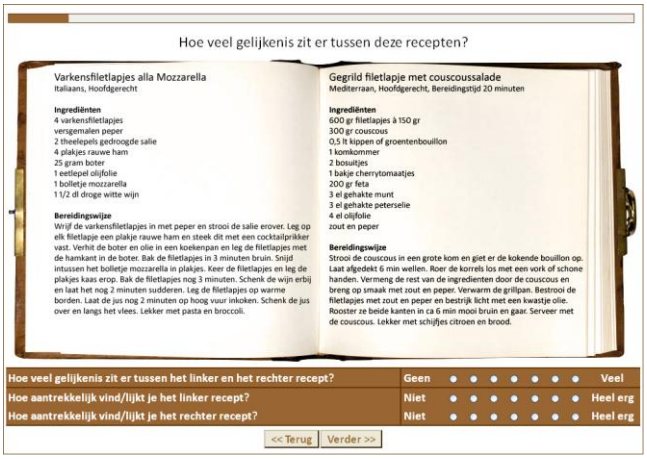

Figure 5. Example Layout page of web questionnaire.

To investigate whether the selection of recipes might have had a spurious effect on the difference between the two similarity measures, we tested whether there is a significant difference between the four randomly composed recipe sets. The results of the corresponding ANOVA indicated that the difference in recipe sets had no significant effect on the difference between the two similarity measure methods, $F(3, 133) = 1.641$, $p > .05$, $\omega = .12$.

An effective recommender system should suggest attractive recipes. Therefore, we would like recipes that are perceived to be similar to an attractive anchor recipe, to be found attractive as well. We compared, with an independent t-test, the attractiveness ratings of non-anchor recipes that were perceived as very similar (score > 4) to an attractive (also score > 4) anchor recipe with the attractiveness ratings of the other non-anchor recipes. The results indicated that on average, a recipe is found to be more attractive when it is similar to an attractive (anchor) recipe ($M = 5.22$, $SD = 0.056$) than when it is not ($M = 4.10$, $SD = 0.035$), $t(2738) = 15.150$, $p < .001$, $r = .28$.

Simple regression analysis was used to test if (the difference in) attractiveness *in itself* significantly predicted the similarity between recipes. Our hypothesis was confirmed, as this was indeed the case: $\beta = -0.186$, $p < .01$. This means that, as β is negative, the similarity between two recipes increases when the difference in attractiveness decreases. However, the results of the regression further indicated that the predictor only explained 3.5% of the variance ($R^2 = .035$, $F(1,2738) = 98.206$, $p < .01$).

Despite this general correlation between similarity in attractiveness and overall similarity, analysis surprisingly showed that this does not hold for two recipes that were both indicated as unattractive. Such pairs were assigned significantly *lower* ratings of similarity ($M = 3.44$, $SD = .079$) than the other recipe pairs ($M = 3.78$, $SD = .034$), $t(2738) = -3.945$, $p < .001$, $r = .075$. This phenomenon is very interesting if it could be replicated, however we are currently not able to explain it satisfactorily.

As the difference in attractiveness did significantly predict the similarity between recipes, we further investigated if the difference in attractiveness significantly also predicted the difference between the two similarity measures. Using simple regression analysis, we found that this was the case: $\beta = 0.086$, $p < .01$, but that the predictor only explained 0.7% of the variance ($R^2 = .007$, $F(1,1368) = 10.118$, $p < .01$). As β is positive, this means that the difference between the two similarity measures increases when the difference in attractiveness increases. Therefore, the difference between the results of the two similarity measures is not only caused by the difference between the two measures itself, but also, albeit only slightly, by the attractiveness of the recipes found by those methods.

Finally, we have looked at some recipe triplets in more detail. For seven out of all 40 anchor recipes, the recipe found with the CSM was perceived to be more similar than the recipe found with our measure. The largest average difference (0.60) was found for the anchor recipe *Genovese spaghetti*, which was thought to be more similar to *pasta with garlic* (found with CSM) than to *caponata with conchiglie pasta* (found with RSM). The (33) anchor recipes for which the RSM-found recipes were more similar showed some more extreme differences; the largest one (2.03) was found for the anchor recipe *chicken masala with potatoes and carjang* and *chicken masala with roti* (RSM) and *rice with frankfurter sausage and catjang* (CSM).

We identified two directions for improvements in the recipe similarity measure. First, for some of the features in RSM, the modeling of the features should be readdressed, especially the weights for starch. For some recipes, it is indicated that they can be eaten with bread, although the ingredient list already contains another type of starch (e.g. pasta). High similarity scores are then computed for recipes that share multiple kinds of starch, such as the *Genovese pasta* and the *Caponata with conchiglie pasta*. In such cases, only identifying the main starch in the feature vector would give more reliable results. Second, we also observed high similarity ratings for recipes that share words in the title. This might be explained by the set-up of the experiment. However, as the title is a condense summary of the recipe, shared title content might actually be indicative for similarity of the corresponding recipes. Therefore, our research confirms that the general approach of combining extracted features and title cosine similarity (introduced by Zhang et. al [32]) is likely to be successful.

From the results of the web experiment, we can conclude that we are well capable to identify similar recipes using the approach as described in the previous sections. The challenge is now to identify how such a measure can be applied to recommending recipes in everyday life.

TOWARDS A RECIPE RECOMMENDER SYSTEM

As argued in the introduction, the goal of this work is to derive a recipe similarity measure that supports people to select more healthful meals in their everyday life. Hence, our challenge is not solely to produce a set of attractive novel recipes, but these recipes should also fit the user's daily routine. The challenge of assessing the healthfulness of a recipe is beyond the scope of this paper. Earlier work showed promising results on the extraction of the amount of vegetables in a recipe [12], which is a good indicator for its healthfulness. Alternatively, either recipes from trusted sources (e.g. the British Nutrition Foundation or the American Diabetes Organization) can be included in the collection, or the healthfulness of recipes in some collection can be manually assessed by an expert.

To learn about the applicability of implementing a recipe recommender system using the proposed recipe similarity measure, an at-home user trial was conducted. In this experiment, a small set of participants was invited to actually cook recommended recipes. The goal of this at-home experiment is to identify directions for an effective

intervention based on tailored recipe recommendations. To this end, recommendations were presented for five consecutive working days. For these days, we base our recommendations on the self-reported meals chosen within a previous period. We expect that meals which are similar to the ones eaten on the same day of a previous week are likely to fit into the daily routines of the participants and are well appreciated.

Participants

Six people (four females and two males) were asked to participate. The design of this experiment required participants to be people who cook at least five times per week.

Procedure

The experiment took three weeks to complete. During the first week, the participants were asked to write down what they had prepared (and eaten) for five days (Monday - Friday). The meals were specified using the 55 identified features. At the beginning of the second week, we calculated four recommendations per day for each participant. These recommendations were sent to the participants in time for them to make any necessary preparations (e.g. shopping, planning) at the end of the second week. In the third week, the participants were asked to cook either one out of the four recommendations for five days. For example, the recommendations for the Monday were based on the reported meal for the Monday in the first week of the trial.

For every reported meal, we used our database of 6000 recipes for main meals to compute an ordering by similarity for those recipes. We selected four recipes as recommendations: the first, tenth, 250th and 6000th recipe in the ordering. By presenting these four alternative recipe suggestions, the participants could choose from recipes that widely varied in similarity to the familiar recipe. This wide variety in similarity was intended to provide some insight into how similar a recommendation should be from a familiar recipe.

Results

During the first week, when the participants had to write down what they had prepared, two of the six participants appeared to have eaten the same meal two days in a row. Therefore, we have calculated recommendations for 28 days only. Table 5.1 shows the average results per recommendation on a 7-point Likert scale.

It can be seen that recommendations 1 (the recipes most similar to the recipes that were prepared in the first week) were chosen most of the times, but recommendations 3 (a bit similar) and 4 (totally dissimilar) were also often prepared. Remarkably, recommendations 2 (quite similar) were chosen considerably less. Despite of this, recommendations 2 were not the most rejected recommendations.

In the last column of Table 5.1, we can see that participants indicated fourteen times (i.e. in 50% of the cases) they would never prepare recommendation 4. This is much more than for the other recommendations, which indicates that totally dissimilar recommendations are generally less adequate than more similar ones. The most similar recommendations (1) were only rejected two times, which confirms the conclusion that most similar recommendations are often acceptable ones. The participants indicated that almost all chosen recommendations fitted in with their normal eating pattern. Also, the vast majority of the recommendations – regardless of the ranking – was perceived to be an attractive recipe. For each day, the participants also answered an open question about the three recipes that were not selected. For a vast majority of the cases, the participants stated that multiple recipes were found to be attractive. Often, the participants stated that they intended to prepare a non-selected recipe on another occasion. These results are in line with the outcomes of the study by Svensson et al. [19], who found that users accepted recommendations even though they were based on a limited or absent user profile.

Table 5. Results of the at-home trail.

Recommen-dation	# times prepared	Fits eating patterns	Will prepare more often	Do not want to prepare
1 (1st)	10	5.6	5.5	2
2 (10th)	3	6.3	4.7	9
3 (250th)	8	6.5	6.5	9
4 (6000th)	7	4.9	5.0	14
Total/average	28	6.0	5.6	34

Apart from assessing the attractiveness of the recipes, the participants were also asked whether the selected recipes meet fit with their desire for more variation. The results showed that recommendations 3 not also fitted well in the participants' normal eating pattern (Table 5), but participants also reported that these suggestions provide a satisfying amount of variation to their normal recipes (i.e. on average 6.00 on a 7-point Likert scale). Participants indicated that especially recommendations 1, but also recommendations 2, provided variations only to a limited extent (5.5 and 5.0 respectively). It is noticeable that these ratings were only provided for the selected recipes. For recommendations 1 and 2 that were not selected, some participants explicitly stated that those recipes were too similar to known ones.

The participants were asked to indicate the reasons for choosing the recipes. Taste, the expected attractiveness of the meals, the ingredients in stock and the fast preparation time were mentioned as reasons to select a recipe. Reasons not to choose a recipe were similarity to the meal they had

the evening before *("I already had pasta yesterday"*) and the lack of availability of ingredients for a more preferred recipe.

Similarity to a known meal was mentioned by two participants as a reason to select a recommended recipe. However, three participants explicitly mentioned similarity to known recipes as a reason not to choose a particular recommendation. These findings suggest that the desired extent of similarity may differ by person or by occasion. In future work, we plan to analyze people's current level of variation in meals. Such an analysis might be used to profile home cooks [10] and predict their desired level of similarity to known meals when trying out new recipes.

From the results, we can conclude that there was at least one good recommendation for every day. This also reveals an important limitation of the set-up of this experiment: as participants could always find a recipe to their liking, ratings were biased towards the positive.

CONCLUSIONS AND FUTURE WORK

In this work, we have derived a measure which models the perceived similarity between recipes for a Dutch audience. To the best of our knowledge, this measure is first in the field of recipe recommendation to be developed by taking a user-centered approach. Important features were identified and were extracted from the recipe texts. Based on these feature vectors, a weighted similarity measure between recipes is presented. This similarity measure was validated in a large-scale web experiment and used in an at-home trial. The similarity measure can be used to promote new recipes that fit into people's lifestyle. Moreover, by assessing dissimilarities between recipes, a varied diet can be offered. When recipes are annotated with a health indicator, the presented strategy may be used to promote healthier alternatives to known meals.

The at-home trial showed that the similarity measure can be used to suggest attractive recipes. However, similarity to a known recipe is not the only criterion for an effective recommendation. Similarity to other meals on previous days, time available, ingredients in stock and the weather and season also showed to be factors that determine the effectiveness of a recommendation. A conversational recommender system, where recipes are presented based on user input may address these requirements. The similarity measure algorithm can be used to select recipes that are similar to familiar ones, but dissimilar to other recipes chosen in that week. In future work, a longer at-home trial with an adaptive system will gain more insights in the usage of recommended recipes.

Alternatively, the approach presented in this work can be combined with the method developed by Freyne and Berkovsky [11]. Taking a user-centered recipe model, the ratings for recipes can be propagated to the ratings for the identified recipe features.

As recipes and food choice are regionally and culturally dependent, the recipe similarity measure cannot be directly used for a different audience. Repeating the card-sorting experiment may reveal different recipe characteristics that are important to other audiences. Future work may show whether such characteristics can also be directly extracted from the recipe texts.

An open problem is the user-friendly assessment of familiar recipes that may not be included in the recipe collection. Currently, these recipes are characterized using the 55 identified features. For a pleasurable application, a more playful method is required to assess current meal choices.

ACKNOWLEDGMENTS

REFERENCES

1. A. Beltran, K. Knight Sepulveda, K. Watson, T. Baranowski, J. Baranowski, N. Islam, and M. Missaghian. Diverse food items are similarly categorized by 8- to 13-year-old children. Journal of Nutrition Education and Behavior, 40(3):149 - 159, 2008.
2. S. Borra, L. Kelly, M. Tuttle, and K. Neville. Developing actionable dietary guidance messages: Dietary fat as a case study. J. of the Am. Diet. Ass., 101:679 - 684, 2001.
3. L. Bouwman, H. te Molder, M. M. Koelen, and C. M. van Woerkuma. I eat healthfully but i am not a freak. Appetite, in press, available online, 2009.
4. C.E. Blake, C.A. Bisogni, J. Sobal, C.M. Devine, and M. Jastran. Classifying foods in contexts: How adults categorize foods for different eating settings. Appetite, 49(2):500 – 510, 2007.
5. J. Brug, A. Oenema, and M. Campbell. Past, present, and future of computer-tailored nutrition education. Am J Clin Nutr 2003;77(suppl):1028S–34S, 2003.
6. J. Burg, I. Steenhuis, P. van Assema, H. de Vries. The Impact of a Computer-Tailored Nutrition Intervention. Preventive Medicine 25:236 – 242, 1996.
7. Canadian Medical Association. 9th Annual National Report Card on Health Care. 2009.
8. J. S. Downs, G. Loewenstein, and J. Wisdom. Strategies for promoting healthier food choices. American Economic Review: Papers and Proceedings. The Psychology of Food Consumption, 99:1 - 6, 2009.
9. A. Drewnowski, S.A. Henderson,A. Driscoll and B.J. Rolls. The Dietary Variety Score: assessing diet quality in healthy young and older adults. J. of Am. Diet Ass., 97(3):266-71, 1997.
10. P. Faye, D. Bremaud, M.D. Daubin, P. Courcoux, A. Giboreau, and H. Nicod. Perceptive free sorting and verbalization tasks with naive subjects: an alternative to descriptive mappings. Food Quality and Preference, 15(7-8):781 – 791, 2004.

11. J. Freyne and S. Berkovsky. Intelligent food planning: personalized recipe recommendation. In IUI '10: Proceeding of the 14th international conference on Intelligent user interfaces, pages 321–324, New York, NY, USA, 2010. ACM.
12. G. Geleijnse, Th. Overbeek, N. van der Veeken and M. Willemsen. Extracting Vegetable Information from Recipes to Facilitate Health-Aware Choices. In proceedings of the Fifth International Conference on Persuasive Technology, Copenhagen, Denmark, 2010.
13. G. Geleijnse, L. Wang, P. Nachtigall, J. Hoonhout and Q. Li. Promoting Tasty Meals to Support Healthful Eating. In proceedings of Wellness Informatics workshop at CHI2010, Atlanta, GA, 2010.
14. K.J. Hammond. CHEF: A model of case-based planning. Proc. AAAI-86, 1986.
15. L. Lähteenmäki and H.C.M. Van Trijp. Hedonic responses, variety-seeking tendency and expressed variety in sandwich choices. Appetite, 24(2):139 – 151, 1995.
16. H.T. Lawless, N. Sheng, and S. Knoops. Multidimensional scaling of sorting data applied to cheese perception. Food Quality and Preference, 6(2):91 – 98, 1995.
17. Maud Lelievre, Sylvie Chollet, Herve Abdi, and Dominique Valentin. What is the validity of the sorting task for describing beers? a study using trained and untrained assessors. Food Quality and Preference, 19(8):697 – 703, 2008. Seventh Rose Marie Pangborn Sensory Science Symposium.
18. P. Nachtigall, G. Geleijnse, J. Hoonhout, A. van Halteren and N. Gros. Toward an intelligent nutritional support system. In proceedings of Wellness Informatics workshop at CHI2010, Atlanta, GA, 2010.
19. Martin Svensson, Kristina Höök, and Rickard C¨oster. Designing and evaluating kalas: A social navigation system for food recipes. ACM Trans. Comput.-Hum. Interact., 12(3):374–400, 2005.
20. R.H. Thaler and C.R. Sustein. Nudge: Improving Decision about Health, Wealth and Happiness. Yale University Press, 2008.
21. S. Thiele and C. Weiss. Consumer demand for food diversity: evidence for Germany. Food Policy 28: 99–115, 2003.
22. Marleen Twigt. Healthy eating & eating varied. Master's thesis, Wageningen University, the Netherlands, 2009.
23. Marlies Wabeke. Creativity in the kitchen. Master's thesis, Wageningen University, the Netherlands, December 2008.
24. National Health Forum. Strategies to increase vegetable and fruit consumption. 1997.
25. Liping Wang, Qing Li, Na Li, Guozhu Dong, and Yu Yang. Substructure similarity measurement in chinese recipes. InWWW'08: Proceeding of the 17th international conference on WorldWideWeb, pages 979–988, New York, NY, USA, 2008. ACM.
26. B. Wansink. Profiling Nutritional Gatekeepers. Food Quality and Preference, 14(4), 289-297, 2003.
27. B. Wansink and P. Chandon. Can "low fat" nutrition labels lead to obesity? Journal of Marketing Research, X(III):605 - 617, 2006.
28. Donna M. Winham. Culturally Tailored Foods and Cardiovascular Disease Prevention. American journal of lifestyle medicine, 3:64, 2009.
29. J. Wisdom, J.S. Downs and G. Loewenstein. Promoting Healthy Choices: Information vs. Convenience. American Economic Journal: Applied Economics, 2(2): 164-178, 2010.
30. World Health Organization. Food and health in Europe: a new basis for action.
31. World Health Organization Regional Publications, European Series, No. 96, 2004.
32. Q. Zhang, R. Hu, B. MacNamee, and S. J. Delany. Back to the future: Knowledge light case base cookery. In ECCBR 2008, The 9th European Conference on Case-Based Reasoning, Workshop Proceedings, pages 239 - 248, Trier, Germany, 2008.