Recipe Recommendation: Accuracy and Reasoning

Jill Freyne, Shlomo Berkovsky and Greg Smith

CSIRO, Tasmanian ICT Center*
GPO Box 1538, Hobart, 7001, Australia
firstname.lastname@csiro.au

Abstract. Food and diet are complex domains for recommender technology but the need for systems which assist users in embarking on and engaging with healthy living programs has never been more real. With the obesity epidemic reaching new levels each day many practitioners are looking to ICT for novel and effective ways to engage and sustain engagement with online solutions. Here we report on a large scale analysis of real user ratings on a large set of recipes in order to judge the applicability and practicality of each. We use traditional content based, collaborative filtering and machine learning algorithms and discuss the trends in the data which reflect user reasoning and decision making strategies.

Key words: Collaborative filtering, content-based, machine learning, recipes, personalization

1 Introduction

The World Health Organisation [1] is predicting that the number of obese adults worldwide will reach 2.3 billion by 2015 and the issue is attracting increased attention. Much of this attention is being paid to online diet management systems, which have been replacing traditional pen-and-paper programs. These systems include informative content and services, which persuade users to alter their behaviour. Due to the popularity of diet monitoring facilities, these systems hold a vast amount of user preference information, which could be harnessed to personalize interactive features and to increase engagement with the system and, in turn, the diet program. One such personalized service, ideally suited to informing diet and lifestyle, is a personalized recipe recommender. This recommender could exploit explicit food ratings, food diary entries, and browsing behaviour to inform its recommendations.

* This research is jointly funded by the Australian Government through the Intelligent Island Program and CSIRO Food and Nutritional Sciences. The Intelligent Island Program is administered by the Tasmanian Department of Economic Development, Tourism, and the Arts. The authors acknowledge Mealopedia.com and Penguin Group (Australia) for permission to use their data.

The domain of food is varied and complex and presents many challenges to the recommender community. The content or *ingredients* of a meal is only one component, which impacts a user's opinion. Others include *cooking methods*, *ingredient costs* and *availability*, *complexity of cooking*, *preparation time*, *nutritional breakdown*, *ingredient combination effects*, as well as *cultural* and *social factors*. Add to this the sheer number of ingredients, the fact that eating often occurs in groups, and that sequencing is crucial, and the complexity of challenge becomes clear.

The paper is structured as follows, the "Recommender Strategies" section provides details the recommendations algorithms implemented in our study which follows in the "Evaluation" section. We conclude the paper with a discussion of our findings and present an outline of future work.

2 Related Work

Initial efforts in addressing these challenges have resulted in systems, such as Chef [5] and Julia [7], which rely hugely on domain knowledge in their recommendation processes. Conversely, fuzzy logic [11] and active learning and knowledge sources techniques have been applied in [14] to generate recipes from ingredient sets without the need for expensive domain knowledge.

Recipe recommender systems have received early attention in the field of case-based reasoning. In the CHEF system [5], 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. [12]. 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.

Zhang et al. [14] also make use of an ingredient representation and in contrast to our approach of treating each ingredient equally they 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 their paper, 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. [24] created a graph-based model of recipes using ingredients and cooking directions. First, they semiautomatically 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 their work, [13] revisit the process of recipe modeling by taking a user-centered approach and do not make any a priori assumptions about the characteristics that determine the perceived similarity, such as ingredients or directions. They derive a measure which models the perceived similarity between recipes by identifying and extracting important features from the recipe texts. Based on these feature vectors, a weighted similarity measure between recipes is determined.

3 Recommender Strategies

This work aims to investigate the characteristics on which people reason about recipes from their actions rather than self reported reasoning. This first step analysis wishes to understand recipes without the context of meal planning or scheduling.

Every recipe has the basic structure of a *Title*, *Ingredients* and *Instructions*. From this information we decipher two measures of complexity, the *number of ingredients* and the *number of steps* required. Finally, we manually annotated each recipe with simple domain knowledge in the form of both a *general cuisine type* and a *specific cuisine type* and a *category* containing options traditionally used to classify recipes based on the meats contained in the recipe. The options for cuisine types and categories are listed in Table **??**.

General Cuisine	Specific Cuisine	Category
African, American, Asian, European, International, Oceania	African, Australian, Chinese, Eastern European, French, German, Greek, Indian, International, Italian, Japanese, Mexican, Middle Eastern, South East Asian, Southern, Spanish, UK&Ireland	beef, pork, chicken, chicken, lamb, fish, veal, vegetables, fruit

African, American,Asian, European,International,UK&Ireland

Figure 1 shows the simple recipe to ingredient relationship strategy adopted in this work. We ignore all cooking processes and combination effects and consider all ingredients to be equally weighted within a recipe. Also, we transfer ratings gathered on recipes equally to all its ingredients, and vice versa, from ingredients to their associated recipes.

We implemented four personalized recommender strategies; two recommender technology approaches and two machine learning approaches suitable for rating prediction. The first is a standard *collaborative filtering* algorithm assigning predictions to recipes based on the weighted ratings of a set of N *neighbours*. Briefly, N neighbours are identified using Pearson's correlation algorithm shown

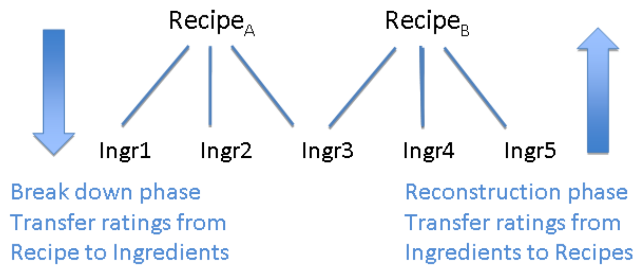

Fig. 1. Recipe - ingredient breakdown and reconstruction

in Equation 1 and predictions for recipes not rated by user u_a are generated using Equation 2.

$$sim(u_a, u_b) = \frac{\sum_{i=1}^{k}(u_{a_i} - \overline{u_a})(u_{b_i} - \overline{u_b})}{\sqrt{\sum_{i=1}^{k}(u_{a_i} - \overline{u_a})^2}\sqrt{\sum_{i=1}^{k}(u_{b_i} - \overline{u_b})^2}} \tag{1}$$

$$pred(u_a, r_t) = \frac{\sum_{n \epsilon N} sim(u_a, u_n) rat(u_n, r_t)}{\sum_{n \epsilon N} sim(u_a, u_n)} \tag{2}$$

The second is a *content-based* algorithm, which breaks down each recipe r_i rated by u_a into ingredients $ingr_1, ..., ingr_x$ (see Figure 1) and assigns the ratings provided by u_a to each ingredient according to Equation 3. The strategy then applies a content-based algorithm shown in Equation 4 to predict a score for the target recipe r_t based on the average of all the scores provided by user u_a on ingredients $ingr_1, ..., ingr_j$ making up r_t.

$$score(u_a, ingredient_i) = \frac{\sum_{l \ s.t. \ ingr_i \epsilon r_l} rat(u_a, r_l)}{l} \tag{3}$$

$$pred(u_a, r_t) = \frac{\sum_{j \epsilon r_t} score(u_a, ingr_j)}{j} \tag{4}$$

We employed a more sophisticated approach to prediction using the open source data mining software Weka [4]. We used the logistical decision tree algorithm M5P [10] to predict scores based on the recipe content and metadata. The algorithm combines traditional decision tress with linear regression functionality at the nodes. Once build the decision tree is built a splitting criterion is used to minimize the intra-subset variation in the class values down each branch stopping if the class values of all instances that reach a node vary only slightly. The tree is then pruned back from each leaf and a smoothing procedure is applied that combines the leaf model prediction with each node along the path back to the root, smoothing it at each of these nodes by combining it with the value predicted by the linear model for that node. The M5P algorithm can be applied all or a subset of the recipe attributes including the presence and absence of ingredients as well as the meta data which accompanies it.

4 Evaluation

We gathered a set of 101557 recipe ratings from 917 users through the Amazon owned online HCI task facilitator Mechanical Turk (www.mturk.com). Online surveys, each containing 36 randomly selected recipes, were posted to the system and users could answer as many surveys as they choose.

4.1 Set-up

Our corpus of 343 recipes was obtained from the CSIRO Total Wellbeing Diet books [8, 9] and from the meal planning website Mealopedia.com (www.mealopedia.com). Each question asked users to report on how much each recipe appealed to them on a 5-Likert scale, spanning from "not at all" to "a lot". 15%(15191) of recipes rated *not at all*, 14% (14425) – *not really*, 20%(19840) – *neutral*, 25% (25593) – *a little*, and 25% (26508) – *a lot*. Each user contributed 110 ratings on average and a ratings matrix of 101557 ratings (33% complete) was gathered.

On average, each recipe was made up of 9.52 ingredients (stdev 2.63) and the average number of recipes that each ingredient was found in was 8.03 (stdev 19.86). Table shows the distribution of recipes over the various options for the

We conducted a number of experiments on the dataset collected using traditional recommender and machine learning approaches to determine algorithm accuracy for recipe rating predictions.

For the traditional recommender approaches of Collaborative Filtering and Content Based recommendations we employed a traditional leave one out off-line analysis, which took each $\{u_i,\ r_t,\ rat(u_i, r_t)\}$ tuple from a user profile and used the algorithms presented in Section 3 to predict the rating $rat(u_i, r_t)$. A set of 20 neighbours were selected only once for each user, based on the entire set of ratings provided. Results for the average MAE for each user for each algorithm are presented.

The machine learning techniques were carried out in Weka [4], each user profile was split into 90% training and 10% test data with 10 iterations carried out, the average MAE over the 1- iterations is presented.

While we acknowledge that these metrics differ we argue that the results presented are sufficient ....

The performance of the recommenders was evaluated using the MAE measure [6] and coverage, i.e. their ability to generate recommendations.

4.2 Results

Figure 3 shows the average MAE of the prediction scores for each initial algorithm presented. As expected, these match earlier results from a similar analysis on an smaller dataset presented in our previous work [3, 2]. The worst performing algorithm is of course the random prediction generator with an MAE of 1.6. The worst performing intelligent algorithm is the M5P (MAE 1.29) which reasons on the ingredient list of each recipe. This is to be expected as the matrix generated from this data is extremely sparse and generally unsuitable for this task. The accuracy of the collaborative filtering (CF) and content based filtering(CB) algorithms are similar with an increase in accuracy of only 0.05 over *CF* obtained by *CB*. A comparison between the *collaborative filtering* algorithm, which treats each recipe as one entity and ignores its ingredients, and the *content-based* algorithm, which considers the ingredients, shows that even the naive break down and reconstruction rules applied here offer significant performance benefits in accuracy.

The best performing algorithms however are the machine learning algorithms which take the recipe meta data into consideration. Looking in particular to the accuracy of the $M5P_all$ and $M5P_meta$ algorithms we note that reasoning on the ingredients of the recipes is less accurate than considering only the metadata. The $M5P_meta$ algorithm is the most accurate with an MAE of 0.99. It is worth noting that we also ran this analysis using a Linear regression algorithm but the results were so similar to that of the M5P algorithm that we omitted the results and discussion for reasons of space.

In terms of the coverage and thus applicability of each approach the *M5P* strategy achieved a 100% coverage rate for each user, whereas the *content-based* strategy obtained 92% coverage and the *collaborative filtering* strategy only 83.8%. Thus the machine learning approach appears to be the best performer overall.

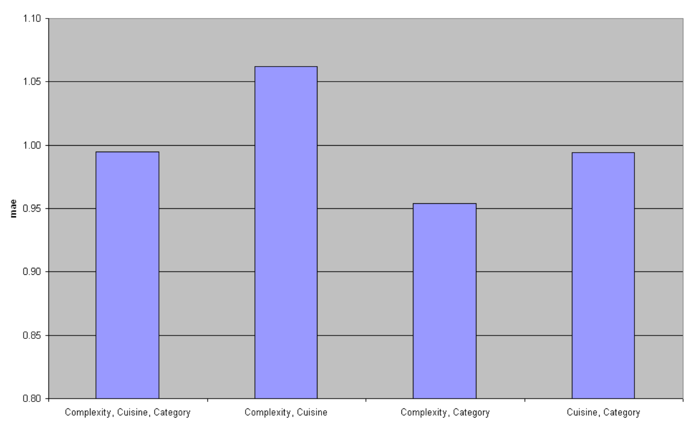

Fig. 2. MAE scores

5 Analysis on reasoning

The purpose of this analysis is broaden our understanding of peoples reasoning on recipes and foods. While knowing which algorithm performs best we needed to do some further investigation on the reasons behind the increased performance. Thus we did further analysis on the metadata and the machine learning algorithm.

As mentioned we use three classes of metadata, *complexity data* which details the number of steps and ingredients in a recipe, *cuisine* data which categorises recipes according to it cuisine type and *category* which categorises recipes according to the main food type included in the recipe.

We ran weka's Attribute Selector analysis on each users profile contained in our dataset. This technique analyses the content of each full profile to determine the suitability of each attribute class in terms of its predictive power for the algorithm selected (in this case M5P).

The selector failed to identify any predictive attributes for 145 user profiles, we hypothesise that these users competed the surveys on Mechanical Turk in a random manner and removed the users from the analysis.

16% of profiles have one predictive attribute, 43% have two, 24% have 3 and 16% have four predictive attributes as seen in Figure 4.

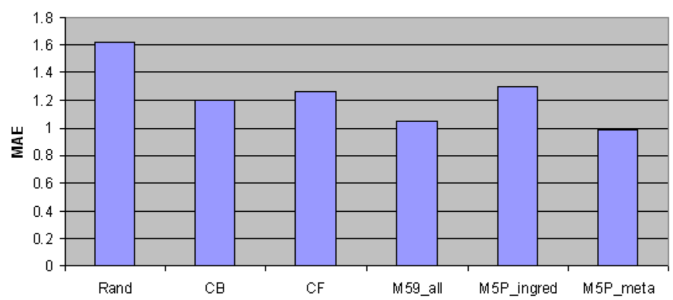

Fig. 3. MAE scores

We hypothesise that different the different number of predictors reflect different thought processes of the individuals who provided ratings. In the case where the category only was a predictor users were providing ratings based on the main ingredients of a recipe, i.e. the meat or poultry content. Where more predictors are present users were predicting on the category in conjunction with the cuisine types giving more or less weight to the predictors based on their preferences.

The *category* attribute is a predictive attribute in 87% of user profiles. In almost all (96.5%) of the cases where only one predictive attribute exists it is the *category* attribute.

43% of users appear to be basing their ratings on two attributes of the recipes.

(under construction ;-))

5.1 Discussion

(under construction ;-)) So does the analysis of the users and their reasoning explain why the CF, CB etc didn't perform as well

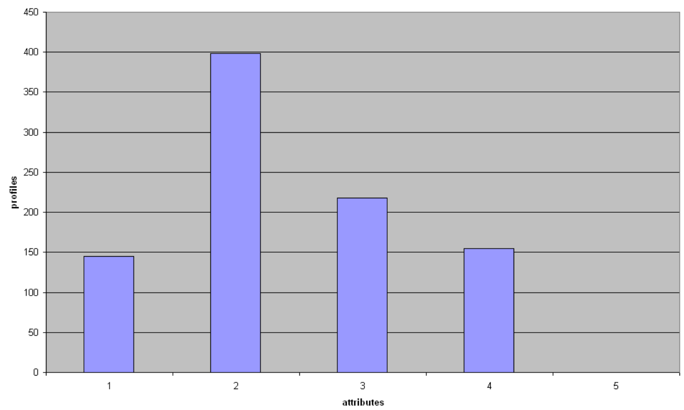

Fig. 4. MAE scores

1. Content based - Reasoning on the ingredients for users to reason on cuisine types is ineffective.... its too low level for the most part

2. Collaborative Filtering - should identify users who reason in the same way ? No?

3. ML - possibly not enough information per profile for the patterns to emerge ? (bit weak)

6 Conclusions and Future Work

In this work we have investigated the applicability of recommender techniques to generate recipe recommendations. We noted reasonable prediction accuracy through traditional content and collaborative filtering algorithms but the highest accuracy obtained by a machine learning technique which exploits decision trees and linear regression to predict highly correlated attributes as predictors of ratings.

Further analysis of prediction algorithm has produced......(under construction ;-))

As noted earlier, there are many factors that influence a user's rating beyond a recipe content. Thus, our future work will focus on extraction of recipe features, such as complexity, time and cooking methods, to examine their impact on user ratings. Furthermore, here we implemented a simplistic idea of what a recipe recommender needs to achieve. We are, however, aware that generating recipe recommendations is a far more complicated task in reality, and we will investigate the issues of group recommendations, where varying social relationships can be at play. In particular, we aim to investigate how family roles and relationships affect compromise and satisfaction with menu plans. Complimentary to this, we need to examine applicability of sequential recommendations. Menu recommendations would not generally be provided in a single shot interaction, but rather users will plan meals over a period of time, such that diversity and satisfaction levels are complex, in particular when groups of users are involved.

References

1. Chronic disease information sheet. http://www.who.int/mediacentre/factsheets/fs311/en/index.html. Accessed June 2010.
2. J. Freyne and S. Berkovsky. Intelligent Food Planning: Personalized Recipe Recommendation. In *International conference on Intelligent User Interfaces (IUI 2010)*, page In Press, 2010.
3. J. Freyne and S. Berkovsky. Recommending Food: Reasoning on Recipes and Ingredients. *User Modeling, Adaptation, and Personalization*, pages 381–386, 2010.
4. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten. The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.
5. K. Hammond. CHEF: A Model of Case-Based Planning. In *Proceedings of the Fifth National Conference on Artificial Intelligence*, volume 1, 1986.
6. J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.*, 22(1):5–53, 2004.
7. T. Hinrichs. Strategies for adaptation and recovery in a design problem solver. In *Proceedings of the Workshop on Case-Based Reasoning*, 1989.
8. M. Noakes and P. Clifton. *The CSIRO Total Wellbeing Diet Book*. Penguin Group, Australia, 2005.
9. M. Noakes and P. Clifton. *The CSIRO Total Wellbeing Diet Book 2*. Penguin Group, Australia, 2006.
10. J. Quinlan. Learning with continuous classes. In *Proceedings of the 5th Australian joint conference on artificial intelligence*, pages 343–348. Citeseer, 1992.
11. J. Sobecki, E. Babiak, and M. Slanina. Application of hybrid recommendation in web-based cooking assistant. In *Proceedings of the Conference on Knowledge-Based Intelligent Information and Engineering Systems*, 2006.
12. M. Svensson, K. Höök, J. Laaksolahti, and A. Waern. Social navigation of food recipes. In *CHI '01: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 341–348, New York, NY, USA, 2001. ACM.
13. Y. van Pinxteren, G. Geleijnse, and P. Kamsteeg. Deriving a recipe similarity measure for recommending healthful meals. In *Proceedings of the 2011 International Conference on Intelligent User Interfaces, IUI 2011*, pages 105–114, 2011.
14. Q. Zhang, R. Hu, B. Namee, and S. Delany. Back to the future: Knowledge light case base cookery. Technical report, Technical report, Dublin Institute of Technology, 2008.