Learning user tastes: a first step to generating healthy meal plans?

Morgan Harvey
Computer Science (i8)
Uni of Erlangen-Nuremberg
91058 Erlangen, Germany
morgan.harvey@cs.fau.de

Bernd Ludwig
Institute for Information and
Media, Language and Culture
University of Regensburg
93053 Regensburg, Germany
bernd.ludwig@ur.de

David Elsweiler
Institute for Information and
Media, Language and Culture
University of Regensburg
93053 Regensburg, Germany
david@elsweiler.co.uk

ABSTRACT

Poor nutrition is fast becoming one of the major causes of ill-health and death in the western world. It is caused by a variety of factors including lack of nutritional understanding leading to poor choices being made when selecting which dishes to cook and eat. We wish to build systems which can recommend nutritious meal plans to users, however a crucial pre-requisite is to be able to recommend dishes that people will like. In this work we investigate key factors contributing to how recipes are rated by analysing the results of a long-term study (n=123 users) in order to understand how best to approach the recommendation problem. In doing so we identify a number of important contextual factors which can influence the choice of rating and suggest how these might be exploited to build more accurate recipe recommender systems. We see this as a crucial first step in a healthy meal recommender. We conclude by summarising our thoughts on how we will combine recommended recipes into meal plans based on nutritional guidelines.

1. INTRODUCTION AND MOTIVATION

In the modern developed world people have the luxury of an abundance of choice with regard to the food they eat. While huge choice offers many advantages, making the decision of what to eat is not always straightforward, is influenced by several personal and social factors [11] and can be complex to the point of being overwhelming [15].

The evidence suggests that many people are making poor dietary choices with stark consequences for their health and well-being. Societal problems such as obesity [19], diabetes [18] and hypertension [14] are all becoming more prevalent, and these conditions are strongly linked to poor dietary habits. The nutritional science literature indicates that these kinds of conditions can be prevented and sometimes even reversed through positive nutritional change [12]. Two issues, though, are that people are generally poor at judging the healthiness of their own diet [8] and even if they do recognise a problem, they lack the requisite nutritional understanding to implement positive dietary changes [4].

Therefore many people could benefit from assistance that allows them to strike a balance between a diet that is healthy and will keep them well and one that is appealing and they will want to eat. After all, it is no good providing users with healthy diet plans if they do not cook and eat the dishes therein, but instead choose unhealthy meals which are more appealing to them.

We believe this is a problem for which recommender systems are ideally suited. If systems can predict dishes that the user would actually *like to eat*, this could be combined within a system modelling expert nutritional knowledge to provide meal recommendations that are both healthy and nutritious, but also appealing. Furthermore complete meal plans for individual users corresponding to nutritional guidelines given by experts could be generated algorithmically which would suit the user's personal tastes. In this paper we work towards these goals via the following main contributions:

- We collect recipe ratings data in context, in a naturalistic setting over a relatively long time period
- Users not only provide ratings data, but specify the reasons behind their rating (i.e. the content and contextual features that led them to rate in this way)
- We analyse the collected data to determine which factors might help us to better understand a user's preferences
- We discuss how these factors could be utilised to build systems which combine recipes into complete meal plans and the challenges this may present

These contributions all relate to the first aim of our work, that is, to better predict which recipes appeal to a given users and are therefore likely to prepare and eat. We conclude the paper by outlining our plans for future work, summarising some ideas on how we may combine recipe recommendations into sensible meal plans.

Paper presented at the Workshop on Recommendation Technologies for Lifestyle Change 2012, in conjunction with the 6th ACM conference on Recommender Systems.

Lifestyle @RecSys'12, September 13, 2012, Dublin, Ireland

2. RELATED WORK

The task of understanding user preferences and suggesting appropriate recipes from a collection can be seen as a novel variant of the well-researched recommender system problem [13, 7]. Although food recommendation is not a frequently studied domain, there is a small body of appropriate related work. Early attempts to design automated systems to plan or recommend meals include CHEF [5] and JULIA [6]. Both of these systems utilise case-based planning to plan a meal to satisfy multiple, interacting constraints. [16] presented a hybrid recommender using fuzzy reasoning to recommend recipes; [9] recommended new food products to supermarket customers, and [17] proposed a system that recommends food items based on recipes recommended to groups of users, clustered by labels.

More recent efforts have tried to better understand the user's tastes and improve recipe recommendations by breaking recipes down into individual ingredients. Freyne and her colleagues [1, 2, 3] demonstrate that this approach works well, with clear improvements over standard collaborative filtering approaches. We wish to build on the success of this work to explore if other content and contextual factors influence the ratings that people assign to recommended recipes. It is our hypothesis that the process of rating a recipe is complex and several factors will combine to determine the rating assigned, beyond purely the user's tastes and that these tastes must be carefully modelled. Both negative and positive ratings could be taken into account, for example: the user may really dislike tomatoes so all recipes with this ingredient might be poorly rated.

Furthermore, not just the existence or absence of explicit ingredients in a recipe but also combination of those ingredients could be important, as could the complexity of the recipe and how long it might take to prepare. Other factors such as how well the preparation steps are described and perhaps the nutritional properties of the dish and the availability of ingredients could have a bearing on the user's opinion of the recommendation. We believe that by building recommender algorithms that incorporate or exploit these kinds of aspects we will be better able to accurately predict ratings. However we also believe that it is vitally important that such factors can be automatically ascertained from ratings data rather than replying on the users themselves. By doing so users can be left to focus on the task of rating recipes and the amount of potentially misleading bias can be minimised. Below we describe how data was collected and analysed to understand how content and contextual factors may influence the way a recipe is rated.

3. DATA COLLECTION

To collect data we developed a simple food recommender system, which selected recipes from a pool of 912 Internet-sourced recipes. This number was chosen as we believe it represents a good balance providing a sufficient variety of dishes from which we may later be able to derive plans whilst, at the same time, being small enough that the resulting ratings matrix will not be too sparse. Users were given a personalised URL and when this was accessed, they were presented with a recipe, selected at random from a list filtered to match a very basic profile. For example, users who specified being vegetarian were only recommended recipes with meta-data indicating no meat; lactose intolerant users were not suggested recipes with milk, etc. Users were not made aware of the random nature of these "recommendations" and were under the impression that the choices were tailored to them. The web page invites the user to provide a rating for the recipe in context i.e. either as a main meal or breakfast for the following day, with recipe meta-data being used to determine which meals should be recommended for which time period. This is important because, in contrast to previous data collection methods, the user is not only rating the recipe with respect to how appealing it is, but also how suitable the recipe is given a specific context. Approximately 3 main meals were recommended for every recommended breakfast.

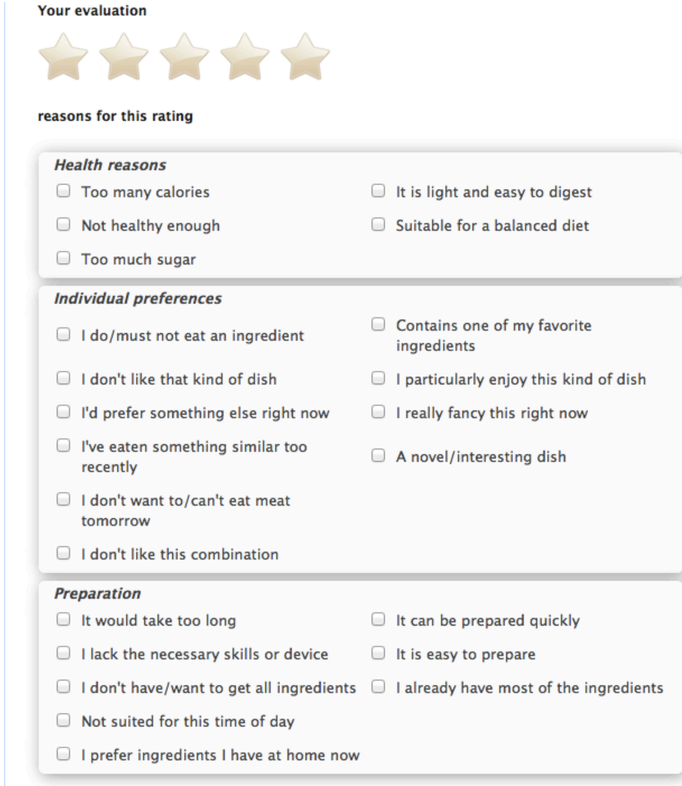

Figure 1: Screenshot of part of the user interface

In addition to collecting ratings, the web interface offered the users the chance to explain their ratings by clicking appropriate check boxes representing different reasons. These check boxes were grouped into reasons to do with personal preferences, reasons related to the healthiness of the recipe and reasons related to the preparation of the recipe – see Figure 1. Reasons contributing positively to the ratings were shown on the right-hand side of the screen and negative reasons to the left. The listed explanations were generated through a small user study, whereby 11 users rated recipes and explained their decisions in the context of an interview. The web interface also provided a free-text box for reasons not covered by the checkboxes, however this was only very infrequently used. We did not record any information regarding whether or not the recipe was later cooked or eaten. We were concerned simply by how appealing the recipe was to the user in the occurring context.

After publicising the system on the Internet, through mailing lists and twitter, 123 users from 4 countries provided 3672 ratings over a period of 9 months. The user population grew organically over time with some users only using the system actively for a few weeks and others for longer periods - the kind of behaviour you would expect with a real system. We argue that although this is a relatively small

Rating	0	1	2	3	4	5
Count	61	818	609	822	828	534
%	1.66	22.22	16.54	22.32	22.76	14.5

Table 1: Breakdown of ratings

and sparse data set, it is an improvement on previous recipe ratings data collection methods, which have used mechanical turk (where there are no validity controls) [1, 3] and surveys where participants rate large numbers of recipes or ingredients in a single session [2]. While surveys can offer the chance to collect data on general user preferences in short time periods, they cannot account for factors, such as food availability, preparation and cooking time, previously eaten meals etc., that would influence ratings if a recipe recommender was to be used in the wild.

Our dataset also differs from previous work in terms of matrix density. The number of ratings per user follows a Zipfian distribution (median = 7, mean = 29.93 max = 395 min =1; 18 users have 1, 52 have 10+). Whereas previous food recommender papers report user - ratings densities of between 22% and 35% [1, 2, 3], our dataset exhibits a user-rating density of 3.28%, which we believe to be much more realistic and more in line with standard recommender systems collections such as movielens and netflix. In terms of ratings per recipe, our collection has a median 3 ratings per recipe (mean = 4.04, max=14, min=2). Table 3 shows the breakdown of ratings (ratings of 0 were discounted as they were marked as not being suitable as a full meal).

Our dataset is, therefore, not only realistic in terms of size, but also a suitable platform for investigation and experimentation as it is both sparse and variant in terms of ratings (sd = 1.41).

4. EXPLORATORY ANALYSIS

To learn about the decision process undertaken when users rate recipes, as well as the factors that influence this process, we analysed the reasons provided by the users when they rated. The aim here was take inspiration for the development of new and improved recommendation models. Figure 2 shows the frequency with which users indicated that particular reasons had influenced the rating they assigned. This figure demonstrates the complexity of the process with several factors - both context and content related - being indicated as being influential. Given that the focus of this work is to inform the development of recipe recommender systems, we focus primarily on factors which could be determined automatically by a system

The most common reasons for negatively rating a recipe (shaded grey in the figure) were that the recipe contained a particular disliked ingredient, the combination of ingredients did not appeal, or the recipe would take too long to prepare and cook. The most common reasons for rating a recipe positively (shaded white) had to do with ease or quickness of preparation, the type of dish or the recipe being novel or interesting. Health related reasons, such as the recipe containing too many calories, the user not perceiving the recipe as being healthy enough, or positive factors like the recipe being balanced or easily digestible were clicked less often overall. However, further analysis revealed that these were clicked very frequently for a particular subset of users. 16.3% of the recipes rated by users who clicked on health reasons at least once had a click on a health reason.

To help understand the relationships between the clicked factors and between the factors and the submitted rating we trained a number of linear models. The final model contained 23 factors in total with 17 factors which were significant i.e. the coefficient estimate is more than 2 standard errors away from 0. Highly significant factors (all p-value $\ll 0.01$) included the combination of ingredients in the recipe, whether the recipe would be suitable for vegetarians, how well the users felt the recipe fitted their own tastes and if the recipe contained a specific ingredient the user particularly likes. All of these significant indicators point to the content of the recipes (in terms of ingredients) being highly significant factors in the choice of rating and also suggest in many cases that this is dependent on the individual tastes of the users. This endorses the approach of Freyne et al., who tried to model ingredient preferences in their work. Nevertheless, the fact that ingredient factors can have both a positive and negative influence on ratings and that the combination of ingredients can be important, suggests that more complicated models may be able to better exploit ingredient information when calculating predictions.

Other important factors included whether to not the recipe would be easy to prepare and whether it suited the time of day specified (i.e. breakfast or main meal) and if the user already had the necessary ingredients at home. Interestingly, given the importance of how easy the recipe is to prepare was, the perceived time required to cook the recipe was not a significant factor. This highlight the complexity of the decision process and the number of factors - context-related and content related - which influence how a recipe is rated.

A number of factors related to how healthy the user perceived the recipe to be including if the user felt it would be light and easy to digest and if the user felt it was too unhealthy. In general these health factors did not contribute significantly to the predictive power of the linear models for all of the ratings together, however we wanted to understand if they might help predict ratings on a *per-user basis*. We looked at the correlation between calorie and fat content of recipes and the ratings provided by two groups of users, those had clicked on a health related factor once or more (Care-about-Health, n = 53, 2572 ratings), and those who never clicked on a health reason (Don't-Care-About Health, n = 70, 1110 ratings)[1]. Figures 3 and 4 show clear differences between the rating behaviour exhibited in these groups. There is a clear trend that the higher the fat content of recipes (r^2=0.88, p=0.012) or the higher the calorific content (r^2=0.87,p=0.022), the lower users in Care-about-Health group tend to rate the recipe. This trend is not present in the second group. If anything there seems to be a slight tendency toward the reverse trend whereby recipes higher in fat ($r^2 = 0.230$,$p = 0.643$) and calories ($r^2 = 0.73$, $p = 0.064$) tend to be assigned a higher rating. This observation suggests that accounting for nutritional factors will allow more accurate recommendations to be generated.

To summarise, these analyses of the collected data demonstrate the complexity of deciding how suitable a recipe will be to cooked in the near future. The results also hint that several factors could be exploited in recommendation algorithms for recipe recommendations.

[1] Nutritional content of recipes was calculated using the system as described in [10].

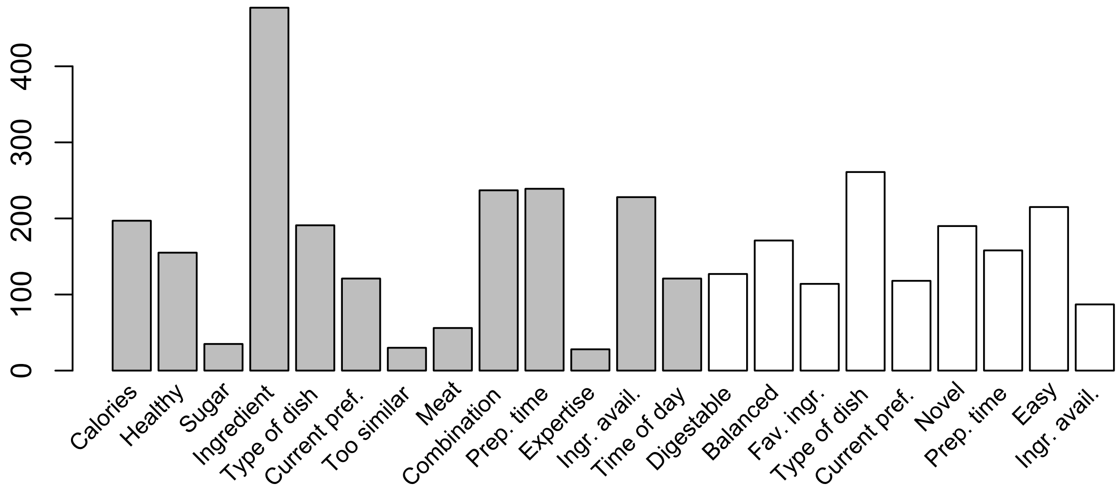

Figure 2: Reasons given for ratings

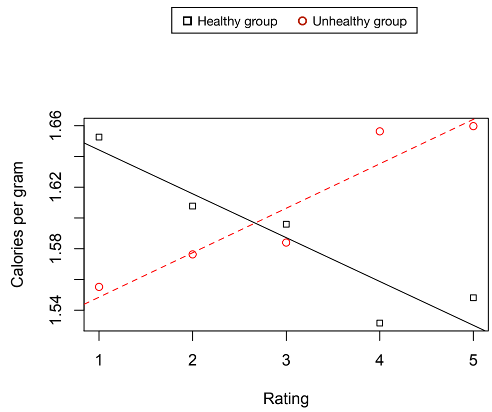

Figure 3: Influence of Calorific Content on Ratings

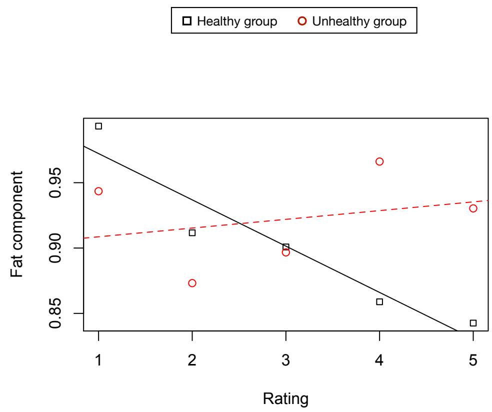

Figure 4: Influence of Fat Content on Ratings

5. BUILDING ON THESE RESULTS

In the previous section we uncovered several patterns in the data indicating that building recommendation algorithms able to account for specific content or contextual features may enable more accurate prediction of recipe ratings. Two important open questions are 1) how can we derive these contextual variables in real-life settings without asking the user to explicitly define their context? And 2) how can we best incorporate such features into recommendation models? We outline some of our thoughts on these points below:

The reasons given by users in our study and the corresponding ratings suggest the ingredients contained within a recipe are very important to the rating process. This finding endorses the approach of Freyne and her colleagues. However, it is clear from our data that ingredients can have either a positive or a negative influence on the rating. For example, if the user likes tomatoes and a recipe contains this ingredient it would be a reason for a high rating. On the other hand, however, if a user does not like tomatoes, our data shows this will negatively affect the recipe rating. Previous recommender algorithms do not account for this negative bias and we believe, based on our results, that including this would improve prediction accuracy. Future recommender models may also account for how important an ingredient is to a dish. For example, imagine a user who does not like tomatoes. For his rating of a recipe where tomato is merely a garnish, this may not have a large influence on the rating. However, if the tomato is a vital ingredient in the recipe e.g. in a tomato soup, then it is more likely to have a large influence.

Another point to consider with respect to ingredients is the coverage of particular ingredients within a collection. For example, Freyne et al.'s algorithm deals with ratings for individual ingredients. This means if egg is rated highly egg-white will be not be treated in the same way. This is exacerbated in our case by the fact that our recipes are web-sourced and may have vocabulary mis-match issues. These kinds of relationships between terms could be identified via instances of nth order co-occurrence. This could be achieved via the use of dimensionality reduction techniques such as singular value decomposition.

Reducing the dimensionality of the feature space would likely have other advantages with respect to dealing how ingredients are combined in a recipe. Our data show that the combination of ingredients can influence the rating applied to a recipe. For example, a user may rate recipes with tomato highly and recipes with pineapple similarly highly on average. However, recipes which combine these ingredients may be given a very low rating. On the other hand, tomato and basil are a combination that work well together and this may have an extra positive influence on the data. Dimensionality reduction techniques, such as SVD or Bayesian Latent Variable models, should implicitly deal with these kinds of patterns.

Our analyses further suggest that including nutritional information in recommendation models should allow more accurate prediction of ratings. We identified two groups of users who behaved very differently based on whether or not they at some point checked that the healthiness of a recipe as an explanation for a rating. The "healthy group" tended to assign a lower rating to recipes higher in calorie and fat content, while the "unhealthy group" displayed, if anything, the opposite predisposition. The group to which a user should be assigned could be obtained explicitly from the user or, preferably, could be learned from ratings data. For example, recipes could be assigned a healthiness score based on nutritional guidelines from health experts and learn which group a user belongs to based on the way they rate recipes with high or low health scores. We acknowledge that the nutrition-aware models may improve performance by offering unhealthy dishes to the users that prefer such dishes and this could be against our long-term goals. We would, however, deal with this issue when combining recipes into meal plans as explained below.

6. CONCLUSIONS AND FUTURE WORK

In this paper we have investigated the decisional process involved in rating recommended recipes. We collected ratings data for recipes and context and statistically analysed the reasons behind assigned ratings. Our future goals in the short term include building on this work to design models that better predict user food preferences using the ideas suggested above. We are continuing to collect data and hope to investigate how performance of models change as the collection size increases.

The presented work represents a single component in a much larger project aimed at building recommender systems that promote healthier dietary choices. In the longer term we plan to move beyond the recommendation of recipes in isolation to recommending dietary plans (7 - 30 days). This involves recommending sequences of recipes under constraints. These constraints will include user preferences of combining recipes and nutritional knowledge, such as the daily recommended intake suggested by the WHO, and user activity patterns. The WHO guidelines provide a means to calculate recommended calorie intake based on a user's profile, as well as a breakdown of the percentage of energy that should come from different types of sources (proteins, fats, carbs, fibre etc.)

One way of modelling this situation is to view it as a graph problem, where the shortest pathes should be computed in a graph where nodes correspond to meals. A week with three meals per day would be represented by a graph with 7 * 3 nodes where edges correspond to dishes (e.g. spaghetti carbonara is an edge from breakfast today to lunch today). A possible cost function could be the distance from the intake estimated from the ingredients and the portion size compared to the recommended daily value. Evaluating the output of such algorithms will be a challenge beyond algorithmics and will involve collaboration with nutritional scientists working on on the project.

Acknowledgements

The authors would like to thank Mario Amrehn and Stefanie Mika for their hard work with the data collection.

7. REFERENCES

[1] J. Freyne and S. Berkovsky. Intelligent food planning: personalized recipe recommendation. In *IUI '10*, pages 321–324, New York, NY, USA, 2010. ACM.

[2] J. Freyne and S. Berkovsky. Intelligent food planning: personalized recipe recommendation. In *IUI '10*, pages 321–324, New York, NY, USA, 2010. ACM.

[3] J. Freyne, S. Berkovsky, and G. Smith. Recipe recommendation: accuracy and reasoning. In *Proc. UMAP*, UMAP'11, pages 99–110, Berlin, Heidelberg, 2011. Springer-Verlag.

[4] J. F. Guthrie, B. M. Derby, and A. S. Levy. *America's Eating Habits: Changes and ConsequencesAgriculture Information Bulletin No. (AIB750)*, pages 243–280. US Department for Agriculture, 1999.

[5] K. Hammond. Chef: A model of case-based planning. In *Proceedings of the National Conference on AI*, 1986.

[6] T. Hinrichs. Strategies for adaptation and recovery in a design problem solver. In *Proceedings of the Workshop on Case-Based Reasoning*, 1989.

[7] D. Jannach, M. Zanker, A. Felfernig, and G. Friedrich. *Recommender systems – An introduction.* Cambridge Univ. Press, 2011.

[8] G. Johansson, A. Wikman, A. M. Ahrn, G. Hallmans, and I. Johansson. Underreporting of energy intake in repeated 24-hour recalls related to gender, age, weight, day of interview, educational level, reported food intake, smoking habits and area of living. *Public Health Nutrition*, 4(4):919–27, 2001.

[9] R. D. Lawrence, G. S. Almasi, V. Kotlyar, M. S. Viveros, and S. S. Duri. Personalization of supermarket product recommendations. *Data Min. Knowl. Discov.*, 5(1-2):11–32, January 2001.

[10] M. Müller, M. Harvey, D. Elsweiler, and S. Mika. Ingredient matching to determine the nutritional properties of internet-sourced recipes. In *Pervasive Health 2012*, 2012.

[11] M. Nestle, R. Wing, L. Birch, L. DiSogra, A. Drewnowski, S. Middleton, M. Sigman-Grant, J. Sobal, M. Winston, and C. Economos. Behavioral

and social influences on food choice. *Nutrition Reviews*, 56(5):50–64, 1998.

[12] D. Ornish, S.E. Brown, J.H. Billings, L.W. Scherwitz, W.T. Armstrong, T.A. Ports, S.M. McLanahan, R.L. Kirkeeide, K.L. Gould, and R.J. Brand. Can lifestyle changes reverse coronary heart disease?: The lifestyle heart trial. *The Lancet*, 336(8708):129 – 133, 1990.

[13] F. Ricci, L. Rokach, B. Shapira, and P.B. Kantor, editors. *Rec. Systems Handbook*. Springer, 2011.

[14] P. Scarborough, P. Bhatnagar, K. Wickramasinghe, K. Smolina, C. Mitchell, and M. Rayner. Coronary heart disease statistics. *British Heart Foundation Health Promotion Research Group Department of Public Health, University of Oxford*, 2010.

[15] B. Scheibehenne, R. Greifeneder, and P. M. Todd. Can there ever be too many options? a meta-analytic review of choice overload. *Journal of Consumer Research*, 37:409–425, 2010.

[16] J. Sobecki, E. Babiak, and M. Słanina. Application of hybrid recommendation in web-based cooking assistant. In *Knowledge-Based Intelligent Info. and Engineering Sys.*, pages 797–804. Springer, 2006.

[17] M. Svensson, J. Laaksolahti, K. Höök, and A. Waern. A recipe based on-line food store. In *5th Int. Conf. on Intelligent User Interfaces*, IUI '00, pages 260–263, New York, NY, USA, 2000. ACM.

[18] Diabetes UK. Reports and statistics on diabetes prevalence.

[19] WHO. World health organization: Chronic disease information sheet http://www.who.int/mediacentre/factsheets/fs311/ (accessed feb 14th 2012).