Ingredient Matching to Determine the Nutritional Properties of Internet-Sourced Recipes

Manuel Müller[1] Morgan Harvey[1] David Elsweiler[2] Stefanie Mika[1]

[1]Chair of Artificial Intelligence

University of Erlangen-Nuremberg, Erlangen-Germany

[2] Institute for Information and Media, Language and Culture

University of Regensburg, Germany

e-mail: ichbin@dermanuel.com, morgan@derharvey.de, david@elsweiler.co.uk, stefanie.mika@cs.fau.de

***Abstract*—To utilise the vast recipe databases on the Internet in intelligent nutritional assistance or recommender systems, it is important to have accurate nutritional data for recipes. Unfortunately, most online recipes have no such data available or have data of suspect quality. In this paper we present a system that automatically calculates the nutritional value of recipes sourced from the Internet. This is a challenging problem for several reasons, including lack of formulaic structure in ingredient descriptions, ingredient synonymy, brand names, and unspecific quantities being assigned. We present a system that exploits linguistic properties of ingredient descriptions and nutritional knowledge modelled as rules to estimate the nutritional content of recipes. We evaluate the system on a large Internet sourced recipe database (23.5k recipes) and examine performance in terms of ability to recognise ingredients and error in nutritional values against values established by human experts. Our results show that our system can match all of the ingredients for 91% of recipes in the collection and generate nutritional values within a 10% error bound from human assessors for calorie, protein and carbohydrate values. We show that the error is less than that between multiple human assessors and also less than the error reported for different standard measures of estimating nutritional intake.**

***Index Terms*—Lifestyle, Health, Prevention, Recommender Systems**

I. Introduction and Motivation

Poor dietary habits are a major cause of global health problems in the modern world. The World Health Organisation (WHO) predicts that the number of obese adults worldwide will reach 2.3 billion by 2015 [1]. In England nearly 1 in 4 adults and over 1 in 10 children aged 2-10 are obese [2], over 5% of the population are registered diabetic [3] and around 1 in 3 adults are hypertensive [4]. Similar statistics can be found for most developed countries and are representative of poor overall health, strongly related to dietary problems. There is a large body of evidence that these problems can be prevented and sometimes even reversed through good nutrition [5], [6].

Two core problems are that people are generally very poor at judging the healthiness of their own diet [7], [8] and even if they do recognise a problem, they lack the requisite knowledge of nutritional principles in order to implement positive dietary changes [9]. There is evidence that as knowledge of nutrition increases eating habits tend to improve [10].

Much of the information people need to improve their diet is freely available. In addition to countless books, magazines and television programmes on the subject, the Internet provides food portals, articles and healthy eating guides that inform about nutrition. Furthermore, databases can be found offering millions of recipes free of charge[1]. These databases provide instructions for meals which in many cases can be prepared simply, quickly and cheaply, with little skill and can be combined to form a healthy and balanced diet. What is lacking, however, is the knowledge, time and motivation required to exploit such resources. In particular people need to understand the nutritional value of individual meals and how they reflect their nutritional needs with respect to their lifestyle.

Technological solutions to help assess and improve diets have been proposed as a solution to this problem. One approach is to design automated systems able to plan or provide meal recommendations for individuals based on their personal nutritional needs, tastes, cooking skills and lifestyle. Early attempts in this direction include CHEF [11] and JULIA [12], both of which utilise case-based planning to plan a meal to satisfy multiple, interacting constraints. More recent efforts have tried to better understand the user's tastes to improve recipe recommendations [13]. Our work aims to build on these initial projects by developing systems that not only recommend recipes according to personal preferences, but combine recipes into dietary plans conforming to WHO nutritional guidelines [14] and user activity profiles derived from sensor technology. We discuss these aims in more detail in Section VIII.

A necessary pre-requisite to building any of the systems described above and implementing them in practical situations is to have appropriate nutritional information for recipes in the database that are available for recommendation. Currently, only small and restrictive datasets can be used, since most recipes available - particularly those obtained from the Internet - have either no associated nutritional data or have data which is partial or inconsistent and often from unreliable sources. In fact, as noted in the literature, there is a general lack of nutrient composition data for dishes and other prepared foods [15].

In this paper we present and evaluate a system that

[1] examples include http://www.cookbooks.com/, http://www.bbc.co.uk/food/ and http://www.chefkoch.de/

automatically calculates the nutritional content of recipes sourced from the Internet. The main contributions of this work can be used in at least two ways. First, the system could be made available as a web service to make accurate caloric and nutritional information more accessible to people cooking at home. Second, it provides a set of annotated recipes that could be used as a dataset for researchers wishing to evaluate techniques for nutritional assistance systems[2].

The remainder of this paper is structured as follows. In Section II, we review appropriate related work, Section IV describes the problem in depth and outlines specific challenges with examples. Section V presents our solution and explains our choice of methods. The evaluation of the system is described in Section VI Parts A and B. We conclude in Section VII with a discussion of what the system provides the research community and outline our plans for future work.

II. Related Work

The gold standard approach for determining the nutritional content of a recipe is to chemically analyse the final cooked dish [16]. Chemical analysis of dishes involves high costs in terms of both time and money. Furthermore this approach cannot be applied in practical situations where a large number of assessments are required in a short period of time (e.g., epidemiological studies, institutional kitchens, private households etc). Considering the many millions of recipes found online, chemical analysis is clearly not a practical solution to the problem.

An alternative is to calculate the nutritional content of meals as part of the cooking process. Smart Kitchen [17] is a pervasive computing kitchen environment that detects and weighs food stuffs and allows the caloric content of the meal to be estimated and monitored by the user as he cooks. Other approaches include using image recognition techniques to analyse pictures of meals consumed. These first detect the main components of meals and then use these to predict the nutritional content based on the results [18]. However, despite work showing that ordinary people are willing to use the approach as part of their everyday lives, the accuracy using current image analysis techniques is very low. Another problem with these approaches is that the user needs to prepare the meal in order to learn its nutritional value.

A further body of research exists focusing on analysing the nutritional content of recipes in a written form. The standard technique is to sum the nutritional value of individual ingredients in an uncooked state [19], [20]. [15] present a number of algorithms which improve on this by accounting for loss of nutritional values through cooking, which will differ based on the nutritional retention of the ingredient and the cooking method. The methods they describe are not easy to implement on large, non-professionally created recipe databases as they rely on the recipe being in a specific format whereby 100% accurate detection of weight, ingredient and cooking method can be achieved. As we will demonstrate, the presentation of the majority of online recipes is such that this is not possible. Nevertheless, previous work shows that simply combining nutrient values for individual ingredients alone can provide acceptably accurate values if the ingredients are selected appropriately [19], [20]. In this paper we work with raw ingredients and focus on the problem of accurately selecting and matching ingredients based on the descriptions given by users when submitting recipes. However, if the ingredient description mentions a specific preparation method e.g. "500g of boiled potatoes" then we use this information to match the ingredient as accurately as possible.

III. Test Collection and Nutrition Database

As a testbed for this system we used recipes obtained from chefkoch.de, a popular German cookery web site with a very large and varied collection of recipes submitted by its users. These users are not food professionals and like the majority of Internet recipe sources there is no editing process. The recipes submitted are in German, which is most suitable for our target users, however we have successfully translated the recipes into English.

In January 2011 we collected a total of 23,500 recipes from the web site containing a total of 39,500 different listed ingredients. The usage of ingredients in recipes follows a power-law distribution with a small number of ingredients featuring very frequently and the vast majority being used far less often. There is a long tail of ingredients that are only used once or twice and these are either exotic or are misspellings. In our database each recipe is represented by a name, a unique ID number, preparation instructions and a list of ingredients, each of which is (normally) composed of both an ingredient description and an appropriate quantity.

In order to ascertain the nutritional content of a recipe, it is necessary to know the nutritional properties of the individual ingredients from which it is made. A prerequisite to building such a system is therefore to have a reliable and extensive table of nutritional values for basic ingredients. Any significant errors in this table would be reflected in values calculated for recipes. There are a large number of freely available tables in the English language, however any translation from English to German for this project would have represented a source of further error. We use the official nutritional table of the German ministry for nutrition, agriculture and consumer protection (Der Bundeslebensmittelschlüssel or BLS) which consists of over 15,000 items and details all required values for each ingredient (energy, fat content, protein, etc)[3]. This table is the largest available German database, is reliably sourced and covers a very broad range of ingredients, including those likely to be used in German cooking which may not be available in English-language databases.

[2] Nutritional data for a collection of online recipes are available from the authors on request

[3] http://www.bls.nvs2.de/

IV. The Problem in Detail

There are two main problems that need to be addressed in order to accurately calculate the nutritional content of a recipe. First, ingredient descriptions in the recipe need to be matched to an appropriate entry in a nutritional database. Second, the quantity of ingredient in the recipe description needs to be converted to a standard scale (in this case, weight in grams). Both of these problems are more challenging than they may appear at first glance. There are several difficulties involved, but these all stem from the fact that users of chefkoch.de (as with the vast majority of Internet recipe databases) are not restricted to using a fixed vocabulary for ingredients and are free to describe the content as they wish. Likewise, users are not forced to describe measurements on a particular self-consistent scale and can choose any description they like. Below we demonstrate the difficulties that can occur with specific examples. First we concentrate on problems relating to ingredient matching. We then shift the focus to converting quantities from the descriptions. While we cannot show all of the challenges involved, we hope the presented examples clearly illustrate the difficulty of the task.

One major challenge relates to ingredient synonymy. Many ingredients have numerous different names, which must be matched to the single term used in the database. For example the word for leek in German can be either “Lauch” or “Porree”, as well as several other regional variants. In Germany, there are huge regional differences in the names used for foodstuffs and this is reflected in the chefkoch collection. This issue also exists in English. Many common examples are a result of the vocabulary differences between British English and American English, for example the salad leaf *eruca sativa* is called variously “rocket”, “roquette”, “rucola” or “arugula”.

A second category of difficulties relates to the level of specificity in recipe descriptions. Some descriptions can be very unspecific, for example in several recipes the ingredient is described as “x fillets of fish”. This is problematic because different kinds of fish can have very different nutritional properties. Other recipes give descriptions such as “4 fillets of white fish”. The system therefore needs to be able to map this description to a particular kind of white fish e.g. haddock. In other examples more specific descriptions are provided e.g. “Fillet of fish (haddock)” , “Filet of fish - haddock” or “haddock filets”. Although the description contains all of the information required to provide an accurate match, the system needs to know that it should match the ingredient named at a particular part of the description and from the examples above, we can see that this position is often variable.

Also relating to the level of specificity in recipes, very common ingredients can have multiple matches in the nutritional table. For example, a search for “tomato” in the BLS returns 100 matches, including different species of tomato e.g. cherry and beef, but also tinned tomatoes, various tomato sauces and soups and even full meals with tomato preparations. These entries have very different nutritional properties and therefore any system would need to be able to accurately choose a single appropriate entry.

As highlighted above, how the ingredient is prepared is also an important factor, with pasta being an illustrative example. For the same weight, cooked pasta has very different nutritional properties to pasta in a raw state because during the cooking process the pasta soaks up water thus changing the weight of the food stuff. In the BLS, there are values for cooked and raw pasta of various types and to make an appropriate match the system must interpret which of these to use. Likewise, smoked meat has different properties to non-smoked and frozen foods have different properties to fresh foods. If this information is available the system should exploit it to maximise accuracy.

Calculating the correct quantity of foodstuff to use can also be problematic. Ideally all amounts would be represented using a common scale (such as weight in grams). However it is often the case with Internet published recipes that other, less specific, weights and measurements are used, e.g. “teaspoon”, “tablespoon” and “cup”. This can often be troublesome because, for example, a cup of liquid weighs more than a cup of fresh herbs. More problematic is that often ingredients are listed with either no amount specified or an unspecific amount, for example “1 tomato”, “a pinch / dash”, “oil for frying” or an ingredient may even be optional in a recipe.

Many of the problems described above can occur in concert. For example in our recipe collection there are at least 4 words for stock (the basis of soups and sauces), these can be of various types (e.g. beef, chicken, fish, vegetable), all of which have different nutritional properties, and the way the amount is generated should be determined by whether it is powdered, concentrated or a liquid.

These examples demonstrate that the lack of controlled vocabulary and syntax in recipe descriptions leads to a number of problems that any system must solve to be useful in practical situations.

V. The Solution

In this section, we first describe the system architecture and explain how the various components work together. We will continue to describe specific components and the work performed to optimise these in detail.

A. System Architecture

An overview of the main components of the nutritional evaluation system and how they function together is shown in Fig 1. The first step is to take the raw description from the source text and separate it into the *amount* and *ingredient description* (1). Both of these parts are then processed separately with the system having components to match ingredients appropriately (Fig 1:right) and to calculate an appropriate weight based on the description text (Fig 1:left). The output from these components is then combined to calculate the nutritional property for each ingredient in the recipe (10). This process is performed for each ingredient description in the recipe and the values summed to calculate the nutritional properties for the complete recipe.

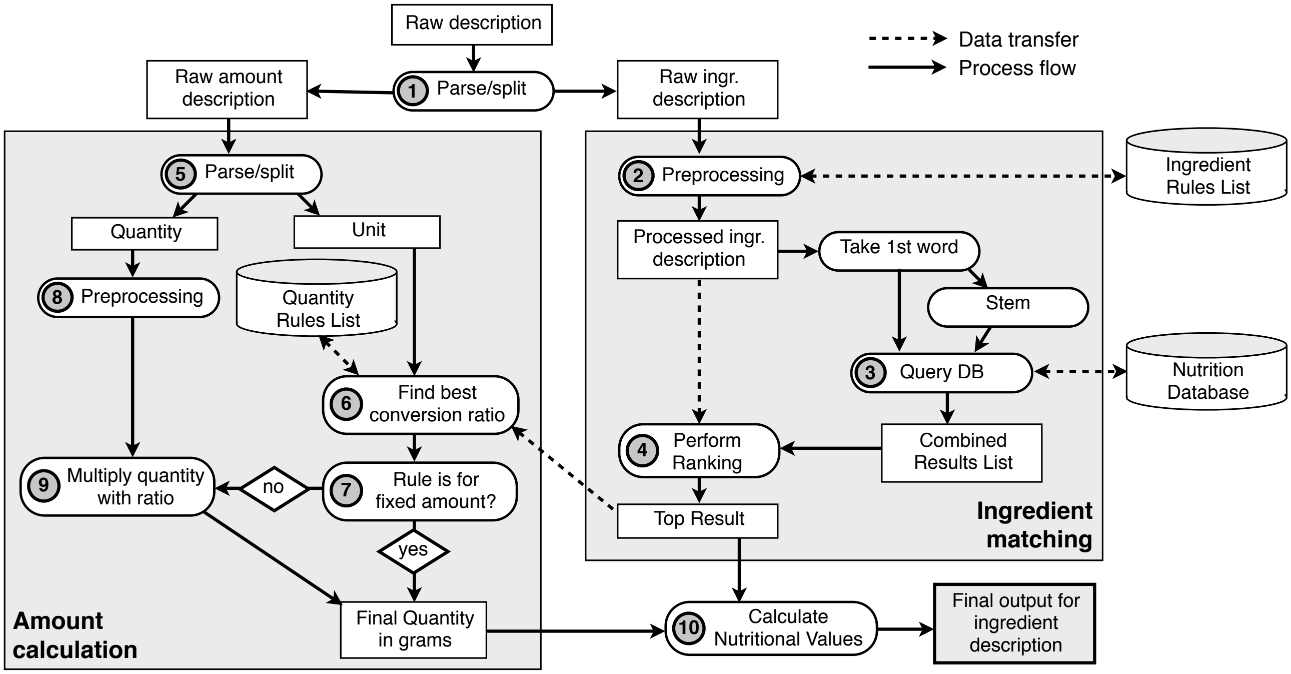

Fig. 1
Diagram of system

To match the ingredient to an appropriate BLS entry, the ingredient description is preprocessed (2) by removing punctuation and converting to lower case. At this stage any appropriate description conversion rules are applied. Rules take the form of “rocket ->ruccola” and “white fish ->haddock”, helping deal with the problem of synonymy and many cases where level of specificity is a problem. A rule-based system is appropriate because the long-tailed distribution of ingredients in the collection means that a small number of rules can cover a large number of problematic situations. Our system currently has 338 ingredient rules, which were created manually for the most common unassigned terms that could not be automatically assigned to an ingredient in the database. The suitability of this rule-based approached is assessed in Part A of Section VI.

The next stage is to match the description or the output of the rule to the database entries (3). The first word in the description is isolated and stemmed and both the original and stemmed versions are used to query the database for any valid matches using the MySQL full-text search function. This combined list of potential matches is then ranked by a weighted ranking model (4), which was trained from a collection of manually provided assignments from human assessors (we explain this modeling process in detail below). The ranking function deals with specificity problems whereby several potential matches have to be reduced to one chosen ingredient with the top-ranked ingredient according to this model being taken as the best assignment and used for the remainder of the process. The influence of this ranking model on system performance is evaluated in Part B of Section VI.

To determine an appropriate weight in grams for the ingredient to be used (Fig 1:left), the raw description of the amount is first split into the quantity and the unit (5). The unit is then checked against a collection of quantity rules (n=198) to determine if a conversion is necessary. Rules take the form of “1 Kg ->1000g”, “1 tablespoon of oil ->15g”, “1 potato ->60g”, “oil for frying ->5g”. Similar to the ingredient rules, the quantity rules list was generated by choosing the most frequently occurring unknown units from our collection and obtaining the correct conversion ratio based on the USDA (United States Department of Agriculture) food database[4]. Again, since the distribution of quantity issues is long-tailed, a small number of rules cover the majority of problematic situations.

To determine which rule should be applied, it is occasionally necessary to know the matched ingredient description. This is indicated by the dashed arrow linking the two components (left and right) in Figure 1. If the chosen rule is for a fixed amount without any specific quantity (7), e.g., “a dash of cream”, then the final quantity in grams is returned. If the amount is not a fixed quantity then the conversion ratio is multiplied by the specified quantity (9) and the final amount is returned. The quantity derived in (5) is processed to deal with fractions e.g. “1/2 Litre of Milk” etc. Once the system has selected a single ingredient from the database and a final amount in grams, then the complete nutritional properties of the item can be calculated and added to the totals for the recipe.

[4] http://ndb.nal.usda.gov/ndb/foods/list

B. Learning to rank

Often database searches (3) can return a large number of results. Since only a single item can be chosen for each ingredient it is necessary to have a system which can rank this list in such a way that the top ranked item has the greatest likelihood of being the most appropriate choice. This problem is closely related to the learning to rank problem which is currently popular in the field of Information Retrieval, where systems seek to improve the quality of ranked lists by observing the clicks made by humans on already presented lists [21].

To learn an appropriate ranking function, we needed a number of data points where an ambiguous query (ingredient description) was given along with the "correct" choice from a list of possible ingredient matches from the database. This can be seen as a two-class classification task where the negative class is poor choices and the positive class is the correct choice. To obtain this data we asked 6 human assessors (researchers at our institution) to evaluate lists of ingredients for ambiguous ingredient descriptions. This process yielded a total of 1,515 positively classed data points to which we added the same number again of negatively classed data points (i.e. incorrectly chosen ingredients). To learn from this data we extracted a number of features from the original ingredient descriptions and the selected ingredients from the database. The choice of these features was driven by both our own intuitions as to what would be useful for differentiating between good and bad choices as well as feedback from the manual ingredient evaluations. In total 16 different features were calculated, however not all of these turned out to be useful.

We constructed a penalised regression model on these features, trained using iteratively re-weighted least squares (IRLS) [22] using 10-fold validation. We use both L1 and L2 regularisation in this case to prevent over fitting and also to determine which features should be discarded, since L1 regularisation has been shown by experiment to be good for this purpose [23]. In addition to the automatic removal of features which are not useful as a result of the aforementioned regularisation we also manually evaluated the usefulness of each feature for regression. For example one of the features which we removed was only positive in 3 data points, which would likely have led to its subsequent parameter in the model being over-fitted.

The final model uses a weighted linear combination of 7 features and outputs values between 1 and -1 indicating the expected relevance of the ingredient to the description and allows ingredients to be ranked in decreasing order of expected utility. The main features used are:

- 3 of the features are counts of how many words or parts of words match between the ingredient description and the ingredient in the database. Since the parts of words comparison is not symmetric this results in 2 separate features; how many times whole words from the ingredient description match partial words in the database description and vice-versa. The full-word match is of course symmetrical. These features ensure that items where larger parts of the descriptions match are given a better ranking. For example if the description says that steak should be "lean" and the database description also has the adjective "lean" in it then it will receive a better ranking score.
- the length of the database description (longer items tend to be more specific, while shorter descriptions are more general).
- the ingredient description does not contain a past-participle e.g. "peeled" or "frozen", items in the database with "raw" in the name are given more weight.
- matching descriptive terms, such as adjectives and past-participles, are given an additional weight in the ranking if they are surrounded by brackets.

The performance of this fitted ranking model and its influence on the system's performance as a whole is evaluated in the following section.

To summarize, our system is built on two main principles. First, available knowledge of food stuffs and nutrition is modelled in the form of rules. Ingredient rules deal mainly with synonymy, while quantity rules help establish the amounts of these food stuffs to account for in nutritional estimations. Second, we exploit linguistic properties of the ingredient descriptions to help the system determine the best ingredient to choose in particular cases. This process has been optimised using a principalled machine learning approach.

VI. Evaluation

We evaluate system performance in two phases. First, we examine the ability of the system to find database entries for ingredient descriptions in the chefkoch collection. This provides an understanding of how widely applicable the system is for the chefkoch collection i.e. the percentage of recipes for which it is appropriate to use the system. Second, to determine the accuracy of matches made, we compare the summed nutritional values for recipes generated by the system - from both the weighted ranking model and an unweighted baseline - against those created manually by human assessors.

A. Matching Ingredients

Without using ingredient rules the system was able to find a match for all ingredients in only 47.2% of recipes and more than 26.4% of recipes had more than 1 ingredient missing. When the rules are used, 91.1% of recipes are matched completely and less than 1% have more than 1 unmatched ingredient. This underlines the benefit of exploiting the long-tailed distribution of ingredients in our rule creation process.

Although these analyses show that our system can identify all ingredients for the vast majority of chefkoch.de recipes, the performance of the system is clearly restricted by the fact that the nutritional table did not contain suitable entries for many common ingredients. The BLS is the most comprehensive German table, however, it lacks entries for common foods, particularly for Asian cooking e.g. coriander leaf, lemongrass and curry powder. The usefulness of our system could certainly be improved by manually adding rules for such foods, exploiting the long-tailed distribution as we did for rules. Food entries could be sourced from reliable, larger English databases such as that provided by the USDA.

B. Nutritional Properties

To establish the accuracy of the system in terms of its ability to match ingredients appropriately, we compared the nutritional output of the system to values generated by human assessors. Note that for this analysis we use the term "calorie(s)" to refer to kilo calories.

To create a "gold standard" with which to compare system performance, we chose a random sample of 50 recipes from the database, ensuring that all of the recipes chosen were main meals and not side dishes, desserts, breakfast dishes or sauces. No effort was made to choose recipes for which the system was able to match ingredients as we wanted to be able to make general statements about system performance. Evaluations were made by a team of 6 researchers led by a nutritional scientist. Each recipe was assigned to 2 evaluators who were asked to manually match each ingredient in the recipe to an appropriate item in the nutritional database.

For each recipe the evaluators were presented with the list of ingredient descriptions and the corresponding weights in their raw form, exactly the same information that is input into the algorithm. They were provided with a basic search tool for finding nutritional values from the database, which functions in the same way as (3) in the system description above, with the exception that no stemming is performed. In addition, assessors were given a list of standard quantity rules, for example "1 tablespoon of oil = 15g". Once an ingredient from the database and a corresponding weight in grams was chosen by the assessor, the resulting nutritional values in terms of energy, protein, fat, carbohydrate and fibre were automatically calculated for the evaluators to input into a evaluation spreadsheet provided. Evaluators were also asked to record the exact ingredient they used from the database for each ingredient description in the recipe and could input any comments regarding the evaluation process that they thought to be relevant.

The evaluations provided by human assessors were in the main very close in terms of nutritional values. In 44.7% of recipes, assessors chose energy values that were within 5% of the mean value between them and in 77% of recipes the difference was less than 25%. The results of this manual analysis, however, illustrate just how challenging this problem can be, as in some cases the agreement between assessors was quite low. In 23% of recipes the error was greater than 25%.

These differences stemmed from both ambiguous ingredient descriptions and unspecific or non-standard amounts. Many discrepancies were caused by each assessor's choice of a specific ingredient when presented with an ambiguous ingredient description. For example when presented with the description "Sausage (smoked)" the two evaluators chose very different types of sausage, resulting in a difference of 580 calories and 53g grams of fat. In another case where the recipe included "2 duck breasts" the evaluators chose very different weights in grams (350g and 150g per breast), causing a difference of nearly 1000 calories and 80 grams of fat. Since it is an infrequently used ingredient, there was no rule for duck breast in the system.

TABLE I
MEDIAN PREDICTION ERRORS PER 100G

	human	**baseline**	**fitted**
Energy (kcal)	10.34	13.8	10.35
Protein (g)	0.54	0.47	0.36
Fat (g)	0.59	1.15	0.69
Carbs (g)	0.73	0.46	0.45
Fibres (g)	0.058	0.058	0.055

TABLE II
MEDIAN PREDICTION ERRORS (% OF HUMAN-EVALUATED MEAN). * INDICATES A SIGNIFICANT DIFFERENCE

	baseline	**fitted**	**gain (%)**
Energy (kcal)	13.03	9.99	30.4*
Protein (g)	7.12	5.02	41.8*
Fat (g)	25.57	12.19	109.8*
Carbs (g)	6.51	5.76	13
Fibres (g)	12.46	10.89	14.4

Tables I and II show the performance of human and the two system versions in terms of error per 100 grams of food and as a percentage of total value for Energy (kcal), Protein (g), Fat (g), Carbs (g) and Fibres (g). For both the baseline and fitted systems the error is calculated as the difference - in calories for energy, otherwise in grams - between the system and the mean of the human evaluators. In this way we take the potential discrepancy between human evaluations into account. The error as a percentage of human-evaluated mean is the difference between the system-derived value and the mean of the human assessments divided by the mean of the human assessments. This provides a unit-agnostic measure of relative system performance which is much easier to use for comparisons than metrics in terms of grams or calories.

We do not report values for vitamins and minerals because these are too dependent on the cooking process, which our system only accounts for if described in the ingredient description.

The main findings of these analyses are as follows:

- The system performed particularly well on calories, proteins and carbohydrates, coming within a 10% error bound for these figures compared to the mean human judgement (see Table II).

- Fat and fibre were more difficult to calculate accurately, with these having a 12.19% and 10.89% error bound respectively.
- The ranking function significantly improves the system performance (Table II:gain). Fat estimations are particularly improved (gain=109.8%). This can be explained by the optimisation of ingredients with specialisations, which often influence fat values e.g. milk (1.5% fat) vs milk (3% fat) or steak vs steak (lean). Differences in the error between the 2 systems for energy, fat and protein are significant according to a dependent 2-group Wilcoxon Signed Rank Test.
- On average, the system with trained ranking (Table I:fitted) is able to provide values closer to one human assessor than the other human assessor can achieve (Table I:human). This underlines both the difficulty of the problem and how well the system performs.

By extracting portion values from chefkoch we can calculate the average error per portion. In doing so we find that for the fitted model the median error per portion is 47.18 calories and 2.96 grams of fat. To put this amount in to context, 47 calories is equal to less than 50g of cooked pasta or 2 teaspoons of sugar and 2.96 grams of fat is around half a teaspoon of butter. Both of these error rates are comfortably within the variance that could be expected when different users interpret and cook the recipe (e.g., through differences in portion size or amount of oil used for frying).

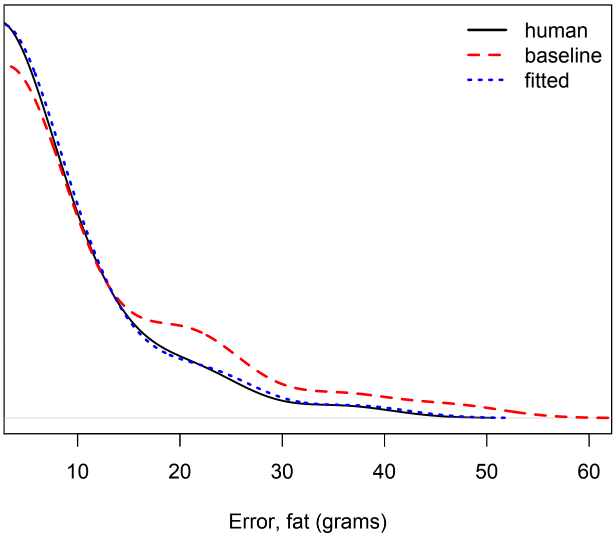

Fig. 2

Error density (fat per portion)

Figure 2 helps provide a better understanding of the system performance. It shows the density of the distribution of errors per portion for fat in grams for the human judgements (human-to-human), as well as for the baseline and fitted systems (to the mean human judgement). These densities are based on a Gaussian kernel smoothed density of the histogram of error rate. All of the nutritional properties follow similar distributions, but the results for fat demonstrate the relationships most clearly.

TABLE III

Pearson correlation coefficients

	Deschamps et al.	**human/fitted**
Energy (kcal)	0.6	0.76
Protein (g)	0.39	0.54
Fat (g)	0.55	0.73
Carbs (g)	0.70	0.6
Fibres (g)	0.51	0.94

The errors for all three have a heavily right-skewed distribution, showing that in the majority of cases the error is very small, however there are a few outlying examples where the error is quite large. These rare cases - where the error is large - are caused by lack of available knowledge, either because the ingredient description is not detailed enough or there is not an appropriate quantity rule for an ingredient. The performance of both the human and system judgements are bounded by this lack of knowledge. Notice that the fitted model has far more of its density in the lower-end of the error scale, showing that it is much more likely to make small errors than the baseline model which has more density in the right-hand tail. The human and fitted model error densities are remarkably similar; in 89% of cases the fitted model makes an error of less than 10g, the humans achieve the same accuracy in 87% of cases and the baseline in only 81%.

Another way to understand the system performance is to compare the error to that reported for methods used to estimate nutritional intake. Deschamps et al. [24] examined the correlation between the nutritional intake values generated using Food Frequency Questionnaires (FFQs) and 24Hr-Recall (24HR) - two methods advocated by the WHO [14] - for a population of 94 adults, adolescents and children. Table III shows the Pearson's correlation scores between the FFQ and 24HR and equivalent scores for our fitted model and mean human assessor value for the core nutritional properties of interest. With the exception of carbohydrate there is less variance between our system and the human assessors than between the two methods of assessing nutritional intake.

VII. Summary

The evaluation sections above show that our system is widely applicable for German recipes sourced from the Internet. The system was able to match all ingredients for 91% of recipes in the chefkoch collection. It was also shown that the system can generate nutritional estimates with low errors for the majority of recipes, that the average error is comfortably within the variance of portion sizes and cooking methods of different users and is more accurate than standard practices for measuring nutritional intake.

When there are large errors in the system, these are normally caused by human assessor interpretation e.g. taking “minced pork” instead of “minced beef” when the recipe calls for “minced meat”. In practise, such problems could be avoided by simply altering the source recipe to use the specific ingredient selected by the system.

Although this work has focused on German language recipes sourced from chefkoch.de, we also have these recipes in English. Furthermore, there is no technological or linguistic reason why the same system architecture could not be used for English recipes. This would involve sourcing another nutritional database, creating a new rulebase and training a new ranking model, but we believe this would function equally well with English recipes.

We argue that this work shows that the collection of chefkoch.de recipes annotated with nutritional data using the system presented here is a suitable starting point for building intelligent nutritional recommender systems for the prevention of future ill-health. We talk about ideas in this direction in the following section.

VIII. Building on this Work

We have a number of research goals that we are actively pursuing. We are working on improving the accuracy of the system by using more sophisticated language analyses to find better features for the ranking model. We are also looking at how recipe instructions can be parsed to understand food preparation methods within recipes, which we know from the literature can improve accuracy.

Beyond estimating the nutritional value of recipes, we have been exploring ways in which these annotated recipe collections can be used. We have been collecting data on user eating preferences and the contextual factors that influence these. For several months 160 users have been rating recipes recommended from the chefkoch.de database and explaining the reasons for their ratings. We want to use our system to help understand if the nutritional content of recipes affects how appealing it is to users e.g. are people more likely to rate calorie rich meals higher than meals low in calories? Is it more important to recommend meals that are easier to prepare? Being able to answer such questions, particularly at the level of individual users, is an important pre-requisite to recommending meal plans that are healthy and that people will actually want to eat.

We are working in collaboration with a nutritionist to develop algorithms that can automatically generate healthy menus for one or several weeks based on the user’s tastes and profile, accounting for features such as novelty and diversity.

References

[1] “World health organization: Chronic disease information sheet http://www.who.int/mediacentre/factsheets/fs311/en/index.html (accessed feb 14th 2012),” .

[2] “Uk department of health (last accessed 14th feb. 2012). http://www.dh.gov.uk/en/publichealth/obesity/index.htm,” .

[3] “Diabetes uk, reports and statistics on diabetes prevalence,” .

[4] 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.

[5] 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*, vol. 336, no. 8708, pp. 129 – 133, 1990.

[6] Diabetes Prevention Program Research Group, “Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin,” *New England J. of Medicine*, vol. 346, no. 6, pp. 393–403, 2002.

[7] E. Brunner, D. Stallone, M. Juneja, S. Bingham, and M. Marmot, “Dietary assessment in whitehall ii: comparison of 7 d diet diary and food-frequency questionnaire and validity against biomarkers,” *British J. of Nutrition*, vol. 86, no. 3, pp. 405–14, 2001.

[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*, vol. 4, no. 4, pp. 919–27, 2001.

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

[10] J. Kolodinsky, J. R. Harvey-Berino, L. Berlin, R. K. Johnson, and T. W. Reynolds, “Knowledge of current dietary guidelines and food choice by college students: better eaters have higher knowledge of dietary guidance.,” *J. of the Am. Dietetic Assoc.*, vol. 107, no. 8, pp. 1409–1413, 2007.

[11] K. Hammond, “Chef: A model of case-based planning,” in *Proceedings of the National Conference on AI*, 1986.

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

[13] J. Freyne and S. Berkovsky, “Intelligent food planning: personalized recipe recommendation,” in *Proceedings of the 15th international conference on Intelligent user interfaces*, New York, NY, USA, 2010, IUI ’10, pp. 321–324, ACM.

[14] Nutrition Joint WHO/FAO Expert Consultation on Diet, the Prevention of Chronic Diseases, and World Health Organization., *Diet, nutrition and the prevention of chronic diseases: report of a Joint WHO/FAO Expert Consultation*, World Health Organization, 2003.

[15] Bognár and Piekarski, “Guidelines for recipe information and calculation of nutrient composition of prepared foods (dishes),” *J. of Food Composition and Analysis Volume*, vol. 13, no. 4, pp. 391–410, August 2000.

[16] Y. Pico, *Chemical Analysis of Food: Techniques and Applications*, Academic Pr, 2012.

[17] P. Chi, J.Chen, H. Chu, and J. Lo, “Enabling calorie-aware cooking in a smart kitchen,,” in *Proceedings of the 3rd international conference on Persuasive Technology*, June 04–06 2008.

[18] K. Kitamura, C. de Silva, T. Yamasaki, and K. Aizawa, “Image processing based approach to food balance analysis for personal food logging,” in *2010 IEEE International Conference on Multimedia and Expo (ICME)*, july 2010, pp. 625 –630.

[19] G. Karg, A. Bognar, and G. Ohmayer, “Nutrient content of composite food: a survey of methods,” in *Proceedings of European Seminar of EOQC Food Section, Budapest*, 1986, pp. 148–179.

[20] P. M. Powers and L. W. Hoover, “Calculating the nutrient composition of recipes with computers,” *J. Am. Diet. Assoc.*, vol. 89, pp. 224–232, 1989.

[21] C. He, C. Wang, and R. Zhong, Y.and Li, “A survey on learning to rank,” in *2008 International Conference on Machine Learning and Cybernetics*. 2008, number July, pp. 1734–1739, Ieee.

[22] P. McCullagh and J. A. Nelder, *Generalized Linear Models (Second Edition)*, Chapman and Hall, 1989.

[23] A. Y. Ng, “Feature selection, l1 vs. l2 regularization, and rotational invariance,” in *21st International Conference on Machine Learning*, 2004.

[24] V. Deschamps, B. de Lauzon-Guillain, L. Lafay, J. Borys, M.A. Charles, and M. Romon, “Reproducibility and relative validity of a food-frequency questionnaire among french adults and adolescents,” *Eur J. Clin Nutr*, vol. 63, no. 2, pp. 282–291, 2007.