INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE

ISSN: 2057-5688

INGREDIENT MATCHING TO DETERMINE THE NUTRITIONAL PROFILE ESTIMATION IN COOKING RECIPES

[1]EDIGA YOGENDRA GOUD, [2]M. ANJANEYULU

[1]PG Scholar, Dept. of MCA, Newton's Institute of Engineering, Guntur, (A.P)

[2]Assistant professor, Dept. of CSE, Newton's Institute of Engineering, Guntur, (A.P)

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. We proposed a scalable approach to estimate the nutritional profile of recipes from their ingredient fraction using a reliable and modern database for nutrient values. Previous research has testified to the performance of string-matching techniques on small data sets. To demonstrate the effectiveness of our system, we tested the proposed approach on a large dataset, RecipeDB, which includes recipes from multiple data sources, including the United States Department of Agriculture standard reference data. Base (USDA SR) is used as a reference. Calculating nutrient profiles. We compared our technique by calculating the average errors in our recipe database (36 calories per serving), which is within the range of errors due to physiological variations.*

Keywords: *Nutrition, Named Entity Recognition, Recipe Dataset, Nutrition Composition Tables, USDA, SR*

I. INTRODUCTION

It is important not to forget the plant part (e.g., leaf, root, stem) when deciding on a comparative diet to estimate nutritional values (Gebhardt, 1992). For example, in the Brassica genus, vitamins from turnips are most appropriate for kohlrabi because both are root vegetables. In contrast, the nutritional values of cabbage, a leafy vegetable, are inappropriate. Vegetable colour is also important when adding a vegetable's carotenoid or nutrient content because carotenoid levels often correlate with green or orange colour. The nutritional A

INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688

content of a dark green vegetable, including broccoli, does not closely resemble the dietary A content of a white vegetable, such as cabbage, even though they are both from the same species. Other factors contributing to nutritional variation in the diets of relatives or the same generation are growing conditions, geographic proximity, plant maturity, processing or preparation, fortification or other ingredients, or meat harvesting (Rand et al., 1991). USDA nutrient database compilers have developed general nutrient profiles for some food agencies. This grouping makes it easy to estimate nutrient values for foods within a group that has not yet been chemically analyzed for certain interesting nutrients. For example, the nutrients and minerals selected for the tropical bottom line acerola, carambola, passion fruit, and chicory were derived from values available for various tropical and subtropical bottom lines.

Estimating the nutritional profile of a cooking recipe is a challenging problem. While there is no dearth of web-based services that provide recipes, their cooking instructions along with ingredient details, pertaining to a wide range of cuisines across the world, their nutritional profiles are not easily available. Here, we propose a Named Entity Recognition(NER)- based strategy for extracting different elements of recipes and to compute the nutritional profile of a recipe by mapping them to their USDA nutritional description. Several methods for the calculation of nutritional values of a cooked meal have been proposed. The most accurate method [1] for this calculation employs chemical analysis. Since this method is applied on the cooked meal, it does not lead to any untoward errors. However, this analysis is not feasible for large datasets of recipes from online resources, since user-uploaded recipes tend to be extremely noisy and without a standard format for storing data. Furthermore, it is not practical to conduct chemical analysis on every recipe, since they may number in hundreds of thousands. Through the

INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688

course of our research, we collected more than 100,000 recipes from one source alone and hence we sought for more scalable methods. An alternative approach is mentioned in [2] where food images are used to calculate calorie contents. Such methods do not provide accurate results suitable for academic research. Since these methods also look for the presence of particular ingredients within food images which are themselves available more accurately in the recipe text, we focus on methods that use the text content itself. The approach we adopted is aligned with the one mentioned in [3] which assumes that the sum total of nutrition of ingredients in a particular recipe can be approximated for the nutritional profile of the recipe. This simplifies our problem statement since we can now calculate the nutritional value of ingredients from nutritional composition tables, and their sum would give us our required nutritional values. It has been observed that more accurate results would be obtained if nutritional yield due to cooking is taken into account, but there is no such consolidated resource for yield values as they differ with ingredient, cooking time and other variable features. Without the knowledge of these variables, it is difficult to estimate the nutritional profile of the recipe with the above method.

II. LITERATURE SURVEY

The gold standard approach for determining the nutritional content of a recipe is to chemically analyse the final cooked dish. 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 is a pervasive computing kitchen

INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688

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 [4].

These first detect the main components of meals and then use these to predict the nutritional content based on the results. 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 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[5].

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. 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

INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688

information to match the ingredient as accurately as possible [6].

III. PROBLEM DEFINATION

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

INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688

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.

IV. PROPOSED SYSTEM

In this paper we present and evaluate a system that 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.

CALCULATING NUTRITIONAL VALUE OF RECIPE

A. Ingredient Data Mining

We utilize the data available from RecipeDB which contains 118,071 recipes from All Recipes and FOOD.com. In order to estimate the nutritional profile of a recipe, we need to obtain all the ingredients used in a recipe and their corresponding quantities, units and/or size and other useful information such as processing state (ground, thawed, etc.), temperature and dryness. Consider the recipe Piroszhki (Little Russian

INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688

Pastries). The Table I shows the outputs of our Named Entity Recognition approach on twelve ingredient phrases. We note that for example, in the table, "1 small onion, finely chopped" contains the entire information that we require to calculate the ingredient's nutritional value, we only need the data in a structured format in order to estimate the nutritional value of the recipe.

We propose a Named Entity Recognition System to train the model to infer the following tags– NAME, STATE, UNIT, QUANTITY, TEMP, DRY/FRESH, SIZE. We manually tagged a corpus of 6612 ingredient phrases and tested the model on a test set of size 2188 ingredient phrases. In order to include ingredient phrases of large diversity in our training and testing set, we utilized Parts of Speech Tagging to form vectors representing each ingredient phrases. A vector representing an ingredient phrase would be defined by the frequency of the tag in the ingredient phrase. We then proceeded to cluster the obtained vectors. The ingredient phrases were chosen for the training and testing set by selecting a subset of ingredient phrases from each cluster. We trained our model using Stanford Named Entity Recognition Model. The model obtained an F1 score of 0.95 on the test set validated by 5-fold cross validation.

Table.1 Examples of Food Description in Usda-Sr Database

S.No	Description
1	Butter, salted
2	Butter, whipped, with salt
3	Butter, without salt
4	Cheese, blue
5	Cheese, cottage, creamed, large or small curd
6	Cheese, mozzarella, whole milk
7	Milk, reduced fat, fluid, 2% milkfat, with added vitamin A and vitamin D
8	Milk, reduced fat, fluid, 2% milkfat, with added nonfat milk solids and vitamin A and vitamin D
9	Milk, reduced fat, fluid, 2% milkfat, protein fortified, with added vitamin A and vitamin D
10	Milk, indian buffalo, fluid
11	Milk shakes, thick chocolate
12	Milk shakes, thick vanilla
13	Yogurt, plain, whole milk, 8 grams protein per 8 ounce
14	Yogurt, vanilla, low fat, 11 grams protein per 8 ounce
15	Egg, whole, raw, fresh
16	Egg, white, raw, fresh
17	Egg, yolk, raw, fresh
18	Apples, raw, with skin
19	Apples, raw, without skin

Closest Description Annotation Using String Similarity Matching

INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688

In order to accurately map ingredient names to food descriptions in the Standard Reference database, we carefully looked for patterns in food description strings that might help us select the best possible description.

(a) It can be observed that the descriptions in the USDASR database are comma-separated terms with a decreasing degree of importance associated with each consequent term. Consider all descriptions from the food description column of Table II. The first term is significant for matching. Hence, Butter, Cheese, Milk, Milk shakes, Yogurt, Egg and Apples occupy the highest priority for finding a match within the ingredient description.

(b) The high priority terms include both singular and plural forms of nouns. They must be lemmatized before matching. For this purpose, we used the NLTK library's WordNet Lemmatizer. Stemmers, although computationally less expensive, were not found to be useful for this purpose because of their high aggression.

(c) The Ingredient Name "Egg whites" best matches with the description "Egg, white, raw, fresh" whereas "Whole eggs" best match "Egg, whole, raw, fresh". The sequence of terms may be different in both the strings being considered. To tackle this, we use a modified form of Jaccard Index as the metric for the similarity between the two strings. The modified Jaccard distance has been explained in (e).

(d) Another observation is that the comma-separated terms in later portions of the food description are more likely to match with the State, Temperature and Freshness of the ingredient. Therefore, we match the whole description along with the State, Temperature and Freshness entities derived from our NER pipeline.

(e) We would like to prioritize the mapping of maximum terms from the ingredient phrase rather than the food description using a vanilla Jaccard Index. This is because a lot of food descriptions include additional details unspecified in the ingredient description. Assume A and B are the

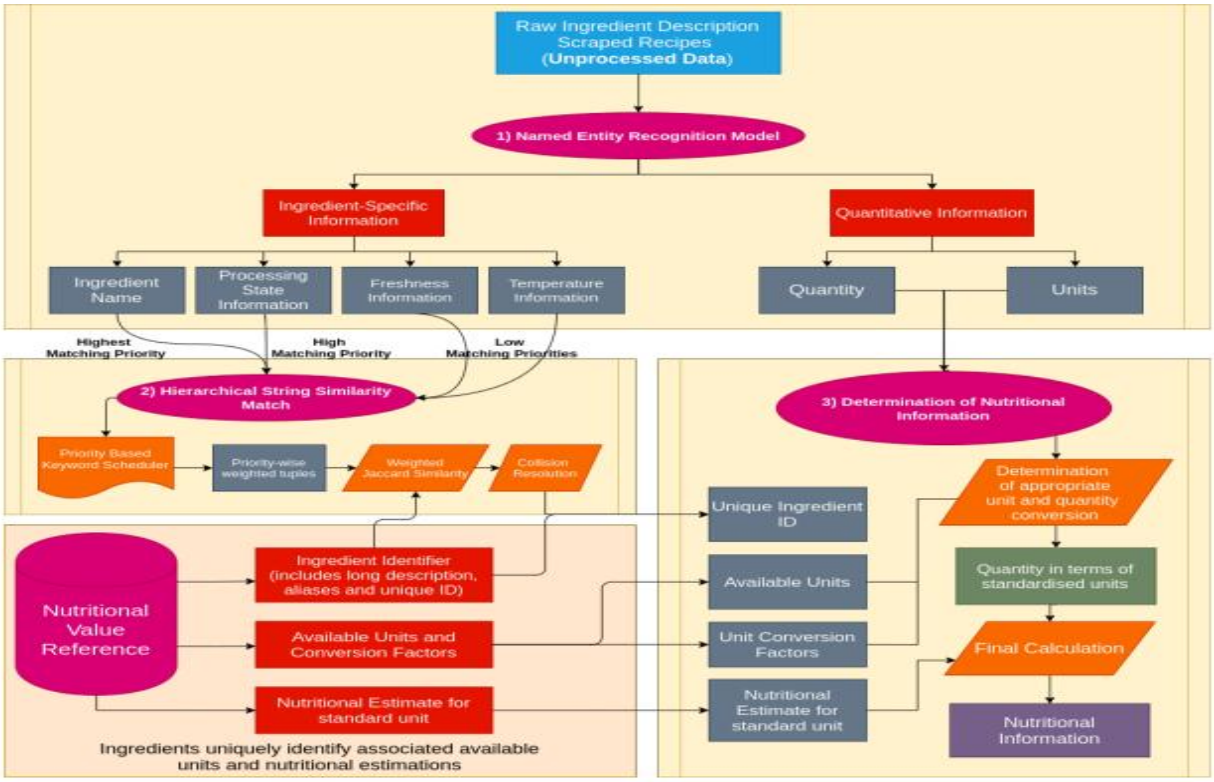
INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688

set of words formed after pre-processing the Ingredient Phrase and Food Description respectively by lemmatization, stop-word removal and uniform casing, and |A| and |B| are the number of words in these sets. Similarly, | A∪B | gives the number of words in the union of the sets of words in strings A and B and | A∩B | gives the number of words in the intersection of the sets of words in strings A and B

For a lot of descriptions, |B| is extremely large, consider food description for serial numbers 7, 8, 9, 13, 14. The denominator in Jaccard distance increases with an increase in |B|. This leads to a bias against large strings. However, it is only essential to match the maximum number of terms from the Ingredient Phrase. So, we use |A| as the denominator for our modified Jaccard Matching Index.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

SYSTEM ARCHITECTURE



INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688

Fig.1 System architecture

C. Units Matching and subsequent Nutrition calculation

Once an ingredient from a recipe has been matched to its corresponding food description in the nutrition database, we match the unit corresponding to this ingredient from our nutritional database. Unfortunately, in this case string matching techniques would not be satisfactory since we have a fixed number of possible units with a fixed format and applying heuristics similar to Section II B might give unwanted results due to incorrect matching of strings. Furthermore, the units provided in the nutritional database may not be enough. For example - our dish requires "1 teaspoon of butter" whereas, in the USDA database, we do not have teaspoon in available units for butter. On top of that some of the units are not clean, e.g., 'pat (1" sq, 1/3" high)' was one of the units in the USDA-SR Database. See Table II Similar problems exist in the units used for the dish. Adding to that, we may have different aliases referencing the same unit in our data, e.g., 'tablespoon' and 'tbsp' refer to the same unit and so do 'pound' and 'lb'. To circumvent these problems, we applied WordNet Lemmatization using NLTK library on all the units present in our recipes and USDA-SR database then took the first word and applied Regular Expression(regex) to obtain a cleaner version containing only alphabets (this helps us to ignore noise and keep relevant part like taking pat out of 'pat (1" sq, 1/3" high)'). Furthermore, standard units were defined for units where aliases were present, for example, tbsp and tablespoon both now represent the standard unit tablespoon. To deal with the case where a unit could not be found, measurement conversion tables were created with detailed conversions between units on the basis of volume using measurements mentioned in [8]. These tables were used to check for the missing units. The tables mention conversions such as '1 cup' is

INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688

equivalent to '16 tbsp' and '48 tbsp' and so on.

Table.2 Ingredient and Unit Relations

ingredient	seq	amount	unit	grams	gram_per_amount
Butter,salted	1	1.0	pat	5.0	5.0
Butter,salted	2	1.0	tbsp	14.2	14.2
Butter,salted	3	1.0	cup	227.0	227.0
Butter,salted	4	1.0	stick	113.0	113.0

V. RESULTS

Using heuristics mentioned in II-B we were able to match 94.49% of the unique ingredients from the recipes, with the rest remaining unmapped from the USDA dataset. To assess the validity of the jaccard matching, the 5000 most frequent ingredients states were manually matched with the USDA dataset, out of which 3580 were deemed to be correct matches, the rest had a better match available in the dataset (accuracy of 71.6%). It is important to note here that USDA has a lot of similar ingredients with little variation as is evident from Table I, so while jaccard similarity does not always give the best match, it almost always gives one of the suitable matches from our database. To further probe how many ingredients along with their units could be mapped to the USDA dataset, we analysed percentage mapping of recipes to their nutritional profile in terms of the percentage of ingredients in a recipe getting mapped to their USDA nutritional profiles (Figure 1). It indicates that the protocol implemented could successfully map a significant proportion of ingredients to their nutritional profiles thereby contributing to the accuracy of the estimated nutritional profiles of recipes. The figure also indicates that the main problem lies in matching the units of ingredients to appropriate units in the USDA dataset, especially when some units are not mentioned in the nutritional database itself. Calorie information from All Recipes was extracted and used as a baseline to evaluate our results. The nutritional profiling of recipes at All Recipes was done by outsourcing it to a reliable third-party. We consider this as the gold standard for the evaluation of our estimated nutritional profiles. We selected data for which we had 100%

INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688

mapping of ingredients with their nutritional values, and had clean, well-defined servings. This resulted in 2482 recipes. This was done because while our recipe dataset has a global coverage, spanning 26 regional cuisines, the sample food composition table that was used mostly contained details of ingredients used in the United States. For e.g., 'garam masala'- a spice used in Indian dishes is not an ingredient present in the dataset. Because of these region-centric ingredients, some ingredients were not mapped. Incorporation of other data as mentioned in Food and Agricultural Organisation of the United Nation5 would help in improving the results.

VI. CONCLUSION

We use NER with Jaccard Similarity and Unit Mapping on a large database containing more than 118,000 recipes to provide accurate estimates of nutritional profiles despite extremely noisy and varied data. We show that the proposed protocol is robust, compatible with any nutritional database, easily replicable and solves one of the foremost problems with dietary analysis and food recommendation systems. We provide the code on Github6 . We would like to highlight that our system provides a good 'estimate' for the nutritional value of food and as nutritional composition tables get updated, our heuristics will give better results without any changes.

RESULTS

[1] Batra, Devansh, Nirav Diwan, Utkarsh Upadhyay, Jushaan Singh Kalra, Tript Sharma et al. "RecipeDB: A Resource for Exploring Recipes.", Preprint.

[2] Al-Maghrabi, Rana. Measuring Food Volume and Nutritional Values from Food Images. Diss. Universit d'Ottawa/University of Ottawa, 2013.

[3] Schakel et al, Procedures for estimating nutrient values for food composition databases,J. Food Compos. Anal, 1997, pp. 102-114

[4] Bognr, A., and J. Piekarski. " Guidelines for recipe information and calculation of nutrient composition of prepared foods (dishes)." Journal of food composition and analysis 13.4, 2000, pp. 391-410.

[5] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling.

INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688

Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics, 2005, pp. 363-370.

[6] Prasadu Peddi (2021), "Deeper Image Segmentation using Lloyd's Algorithm", ZKGINTERNATIONAL, vol 5, issue 2, pp: 1-7.

[7] Niwattanakul, Suphakit, et al." Using of Jaccard coefficient for keywords similarity." Proceedings of the international multiconference of engineers and computer scientists. Vol. 1. No. 6. 2013.

[8] Francis T. Lynch.The Book of Yields: Accuracy in Food Costing and Purchasing, 8th Edition,201

[9] Prasadu Peddi (2016), Comparative study on cloud optimized resource and prediction using machine learning algorithm, ISSN: 2455-6300, volume 1, issue 3, pp: 88-94.

[10] Afreen Bari, Dr. Prasadu Peddi. (2021). Review and Analysis Load Balancing Machine Learning Approach for Cloud Computing Environment.Annals of the Romanian Society for Cell Biology,25(2), 1189–1195.