# POS Tagging-probability Weighted Method for Matching the Internet Recipe Ingredients with Food Composition Data

Tome Eftimov[1,2] and Barbara Koroušič Seljak[1]

[1]*Computer Systems Department, Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia*
[2]*Jo˘zef Stefan International Postgraduate School, Jamova cesta 39, 1000 Ljubljana, Slovenia*

Keywords: Part of Speech Tagging, Probability Model, Information Retrieval, Food Composition Databases, Ingredient Matching.

Abstract: In this paper, we present a new method that can be used for matching recipe ingredients extracted from the Internet to nutritional data from food composition databases (FCDBs). The method uses part of speech tagging (POS tagging) to capture the information from the names of the ingredients and the names of the food analyses from FCDBs. Then, probability weighted model is presented, which takes into account the information from POS tagging to assign the weight on each match and the match with the highest weight is used as the most relevant one and can be used for further analyses. We evaluated our method using a collection of 721 lunch recipes, from which we extracted 1,615 different ingredients and the result showed that our method can match 91.82% of the ingredients with the FCDB.

## 1 INTRODUCTION

It is evidence based that a healthier diet is required to prevent diet-related chronic diseases and to increase the quality of life. However, to assess the quality of a diet, advanced approaches still need to be developed. There is a lot of information about healthier diet and nutrition principles presented in different forms, available in books, magazines, television programs and Internet. But from other side, people are lacking of knowledge about all the nutrition principles and also lack of time and motivation to explore the resources where this kind of information is presented.

A lot of free data sources that contain recipe databases exist and can be used for nutritional assistance or recommendation systems. For this purpose, it is important to have accurate nutritional data for recipes, but most of the recipes have no such data available or have data of suspect quality. The most important is that people need to understand the nutritional value of the individual meals and also how they reflect their nutritional needs with respect to their lifestyle.

In the past, different technological solutions were represented, dealing with problems to assess and improve diets. They used the information from the recipes and food composition data. Food composition databases (FCDBs) provide detailed information on nutritional composition of foods, usually from a particular country. They contain information for a huge number of components including: energy, macronutrients and their components, minerals and vitamins. Food composition data is used for planning diets with specific nutrient composition in clinical practice and for assessment of the nutritional value of the food consumed by individuals and populations (H. Greenfield and D. Southgate, 2003).

Using all this information is useful to generate a system that automatically calculates the nutritional value of the recipe and than the recipe can be used in planing the diet for some individuals or populations. The main problem is that the information on the Internet is incomplete - on the other side FCDBs are lacking of recipes and as chemical analysis is costly, we need to find a way of calculating nutritional values for recipes from the Internet considering food composition data of recipe ingredients. One of the key problems is a lack of structure in the names of the ingredients used in the recipes and a lack of structure in the names of the food analyses from the FCDB. To calculate the nutritional value of the recipe, we need for each ingredient from the recipe to find the perfect or the most relevant ingredient match from the FCDB. For example, we can find "chicken breast, raw" in a recipe, and several food analyses in the



Eftimov, T. and Seljak, B..
POS Tagging-probability Weighted Method for Matching the Internet Recipe Ingredients with Food Composition Data.
In *Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2015) - Volume 1: KDIR*, pages 330-336
ISBN: 978-989-758-158-8

FCDB, which can be "chicken breast, cooked, salted" or "raw chicken breast" or other name that contain "chicken breast". This is the problem from which depends how accurate will be the calculated nutritional value for the recipe using the food composition data presented into the FCDBs.

In this paper, we present an information retrieval method, which is a probability weighted method that enables us to perform a search and find the most relevant match for each recipe ingredient in the FCDB. After having the most relevant match, we can use the data from the FCDB to calculate the nutritional value of the recipe.

In Section II, we review appropriate related work. Section III describes the problem in depth. In Section IV, we present our solution and in Section V, the evaluation and results are presented using real data. Section VI provides the discussion of the results, the benefits of our method and comparison with other approaches presented in the literature. In Section VII, we conclude the paper by discussing the proposed method and our plans for future work.

## 2 RELATED WORK

The task of matching text concepts to an entry in a knowledge base is a very popular one and has been addressed in many ways. In 1988, term-weighting approaches in automatic text retrieval systems were presented, which could be designed on a comparison between the stored text and users' information queries (Salton and Buckley, 1988). Another approach is POS tagging, which means automatic assignment of descriptors, or tags, to input tokens, where the tags are the appropriate grammatical descriptors to words in text. POS taggers can be used for several purposes, and one of them is for text indexing and retrieval, which can benefit from POS information (Schmid, 1994; Tian and Lo, 2015). The task of matching concepts in text has progressed a lot and there are different methods for automatic text retrieval systems. In (Mihalcea and Csomai, 2007), an automatic text annotation system was presented, which combines keyword extraction and word-sense disambiguation to identify relevant links to Wikipedia pages. The system is known as "Wikify", and involves automatically extracting the most important words and phrases in the document (keywords) and identifying for each keyword the appropriate link to a Wikipedia article. Another approach is the entity linking (EL), which is the task of linking name mentions in text with their referent entities in a knowledge base. One method dealing with EL is presented in (Han et al., 2011), which is a graph-based collective EL method, which can model and exploit the global interdependence between different EL decisions. Also, there are approaches that are dealing with automatic ontology based knowledge extraction, for example the Artequakt project presented in (Alani et al., 2003) links a knowledge extraction tool with an ontology to achieve continuous knowledge support and guide information extraction.

Technological solutions have been proposed to improve recipe recommendations. The idea is to design systems that are able to provide meal recommendations for individuals based on their nutritional needs and lifestyle. One approach is presented in (J. Freyne and S. Berkovsky, 2010), which give recommendations of healthy recipes. In order to give recommendations, we need to calculate the nutritional content of a recipe, which can be done using chemical analyses of final cooked dishes (Y.Picó, 2012) or having a system that automatically calculate the nutritional content of a recipe. In (M. Muller et al., 2012), the authors presented a system that automatically calculates the nutritional content of recipes sourced on Internet. To match the ingredient to an appropriate entry from the official nutritional table of the German ministry for nutrition, agriculture and consumer protection, the ingredient name is preprocessed by removing the punctuations and converting to lower case. Because the database search can return numerous results and only a single item can be chosen, they presented a system which can rank the list and the top ranked item need to be used as appropriate match. To learn the ranking function, they treated the problem as two-class classification task where the negative class is poor choices and the positive class is the correct choice. To obtain the data, they asked 6 researches to evaluate manually lists of ingredients for ambitious ingredient descriptions. To learn from the data, they extracted a number of features from the original ingredients name and the selected ingredients from the database. At the end they performed penalised regression model, where the output is between -1 and 1 inciting the expected relevance of the ingredient to the name. Using this method, 91.1% of the recipes they used were matched completely and less than 1% have more than one unmatched ingredient.

## 3 PROBLEM DEFINITION

The problem we want to address is to find the most relevant match for the ingredients used in the recipes using their name and the names of the food analyses that are presented in the FCDBs.

There are several issues that we need to consider when we want to solve this problem. As we said before, there are a lot of online data sources which provide recipes. Some of them allow the users to log into the system and to submit their own recipes, so everyone can use it. The first issue is that people use the human natural language, which is the main vehicle through humans transmit and exchange information, and write the name of the used ingredients in the unstructured form. For example, in different recipes we can find "salt, iodised", "iodised salt" or "salt-iodised". From this, we can conclude that the lack of structured way of representation is presented, and this happens because of the different ways of people expression. Another issue is the ingredient synonymy problem. So we need to match the synonyms to the single term which is used in the databases. Some ingredients can have multiple matches in the food composition database. For example, if we are looking for "salt", we can find "salt", "salt, table", "salt, iodised" and many more. But all these matches may have very different nutritional properties, so we need to chose the most relevant one. Also, a very important factor when we want to calculate the nutritional properties is the preparation method of the ingredient. It is different to have cooked or raw ingredient, for example, "smoked ham" and "non-smoked ham", "chicken breast, raw" and "chicken breast, cooked", because they have different nutritional properties.

All of these issues need to be considered and need to be solved when we want to find the relevant ingredients matching, which can be used to calculate the nutritional value of the recipes.

# 4 POS TAGGING-PROBABILITY WEIGHTED METHOD

One method for ingredient matching is presented in (M. Muller et al., 2012). The method treats the problem as two-class classification problem, which required evaluation by nutrition experts, and after that they use a linear regression model to match the ingredients.

Intend to solve the ingredient matching problem with food composition data, we looked for the existing ontologies in this domain (LIRMM, 2015; Ontology, 2015), and we have found that there are focused on food recipes, ingredients and nutrients, but an information about the structure of the ingredient name is still missing. An ingredient name is represented by noun, and it can be additional explain with the form of the ingredient (adjective) and the cooking process (verb), which are very important and need to be considerate in case when we want to calculate the nutritional value. Have in mind the importance of the nouns, adjectives and verbs presented in the ingredient name, the Part Of Speech tagging (POS tagging) is one technique that can be used for ingredient matching with food composition data (A. Voutilainen, 2003).

Our method is a probability method with which we assign a weight on each matching and we considered the match with the highest weight as the most relevant one. First, for each ingredient from the recipe, we use POS tagging, also called grammatical tagging or word-category disambiguation, to identify the nouns, verbs and adjectives. The nouns carry the most of the information of the name, the adjectives explain the ingredient in most specific form, for example "frozen", "fresh", and the verbs are at the most cases related with the preparation method, for example "cooked", "drained" etc. Then, we search the FCDB for the ingredient with a simple SQL search using the provided nouns from the ingredient name in the recipe. For each found name as a result of the SQL search, we also perform POS tagging to identify the nouns, verbs and adjectives. Next, we define an event (X) which is the similarity between the ingredient name from the recipe and each of the food names that are returned from the SQL search of the FCDB. At the end, the weight we assign to the matching pairs is the probability of the event.

Let $D_1$ be the name of a single ingredient from the recipe, and $D_2$ is the single food name which is a result from the SQL search of the FCDB. Let's define,

$$\begin{aligned} N_i &= \{nouns\ extracted\ from\ D_i\}, \\ V_i &= \{verbs\ extracted\ from\ D_i\}, \\ A_i &= \{adjectives\ extracted\ from\ D_i\}, \end{aligned} \tag{1}$$

where $i = 1,2$.
To find the probability of the similarity between the ingredient name from the recipe and the food name from the FCDB, we present the event as a product of three other events.

$$X = N \cdot V \cdot A, \tag{2}$$

where $N$ is the similarity between the nouns which are in $N_1$ and $N_2$, $V$ is the similarity between the verbs which are in $V_1$ and $V_2$ and $A$ is the similarity between the adjectives which are in $A_1$ and $A_2$.
Because all these events are independent, the probability of the event X can be find as

$$P(X) = P(N) \cdot P(V) \cdot P(A). \tag{3}$$

Now, we need to define the probabilities of each of the events, $N$, $V$ and $A$. Because we want to find the similarity between two sets, it is logical to use the Jaccard index, $J$, which is used in statistic for comparing

the similarity and diversity of sample sets (R. Real and J. M. Vargas, 1996). For this purpose, we use the modification of the Jaccard index in combination with Laplace probability estimate. We do this because in some ingredients description the additional information provided by the adjectives or verbs can be missing, but we can also find the relevant match into the FCDB, so we will have non-zero probabilities. The probabilities of the events can be find as

$$P(N) = \frac{|N_1 \cap N_2| + 1}{|N_1 \cup N_2| + 2} = \frac{J(N_1, N_2) + \frac{1}{|N_1 \cup N_2|}}{1 + \frac{2}{|N_1 \cup N_2|}}$$

$$P(V) = \frac{|V_1 \cap V_2| + 1}{|V_1 \cup V_2| + 2} = \frac{J(V_1, V_2) + \frac{1}{|V_1 \cup V_2|}}{1 + \frac{2}{|V_1 \cup V_2|}}$$

$$P(A) = \frac{|A_1 \cap A_2| + 1}{|A_1 \cup A_2| + 2} = \frac{J(A_1, A_2) + \frac{1}{|A_1 \cup A_2|}}{1 + \frac{2}{|A_1 \cup A_2|}}. \quad (4)$$

We obtained the probability of the event $X$, substituting the relations (4) into the relation (3), which is the weight we assigned to each matching pair and at the end, the pair with the highest weight is the most relevant found match.

Important aspect of the ingredient matching is also pre-processing. First each ingredient name without the difference from where is it, we converted in a lower case letters and also we removed the punctuations. For the nouns, we use lemmatisation to avoid the difference between the singular and the plural form of the noun (J. Plisson et al., 2004). Because, there are names that contain "without skin" and some other "skinless", or " with salt" and "salted", we mapped all of these phrases using rules which we created manually, and are specific for this area. In Figure 1, the architecture of the proposed method is presented.

# 5 EVALUATION AND RESULTS

We performed the evaluation of the method by two experiments. The first experiment is not the proper evaluation of the method, but an illustration of the problem that we are trying to solve, while the second one is the matching between the Internet extracted ingredients and the food composition data.

The data we used for evaluation is a collection of 721 recipes written in English, from which we extracted 1,615 different names of ingredients. We collected it using an HTML parser and a free recipes web site (AllRecipes, 2015). For each of the recipes, we considered only the names of the ingredients, while the quantity-unit pair associated with the ingredient was ignored, as our global goal was to find the ingredients matching.

**Algorithm 1:** POS tagging-probability weighted method.

**for** each ingredient name in recipe **do**
    - set $matching\_pairs = null$
    - set $counter = 1$
    - ingredient name pre-processing
    - extract the sets of nouns $N_1$, verbs $V_1$, and adjectives $A_1$ using POS tagging
    - query the FCDB using the set of provided nouns $N_1$
    **for** each food name from the result of searching the FCDB **do**
        - food name pre-processing
        - extract the sets of nouns $N_2$, verbs $V_2$, and adjectives $A_2$ using POS tagging
        - calculate $P(X) = P(N)P(V)P(A)$
        - $matching\_pairs[counter] = P(X)$
        - $counter = counter + 1$
    **end for**
    - return the most relevant match, $\max(matching\_pairs)$
**end for**

We used the EuroFIR FCDB as our database. EuroFIR AISBL is an international, non-profit Association under the Belgian law (EuroFIR, 2015). Its purpose is to develop, publish and exploit food composition information and to promote international standards to improve data quality, storage and access. EuroFIR presented data model for food composition data management and data interchange. The EuroFIR FCDB contains analyses from several European countries.

We extracted 44,033 English names of foods analyses, which exist in the EuroFIR database. Before we start with the evaluation, we preprocessed the ingredients names from the recipes and the food names from the EuroFIR FCDB. First, we removed the punctuations from them, and then we converted them in lower-case letters.

## 5.1 Experiment 1

The first experiment we made is the ingredients matching for one recipe and it is not the proper evaluation of the method, but an illustration of the problem that we are trying to solve. We used the recipe for "World's Best Lasagna", extracted from (AllRecipes, 2015). The result of the ingredients matching is presented in the Figure 2. Using the information presented in the Figure 2, for the recipe that contains 20 ingredients, we were unable to find match only for

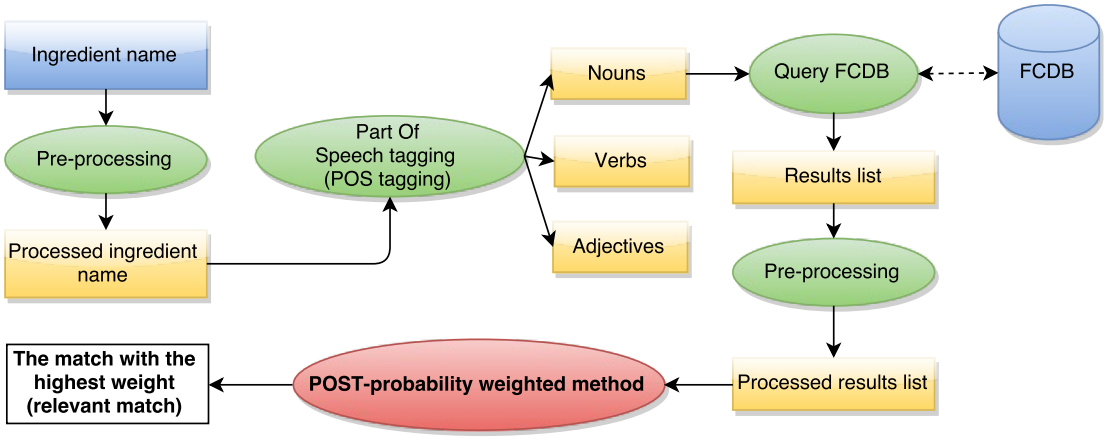


Figure 1: Architecture of the method.

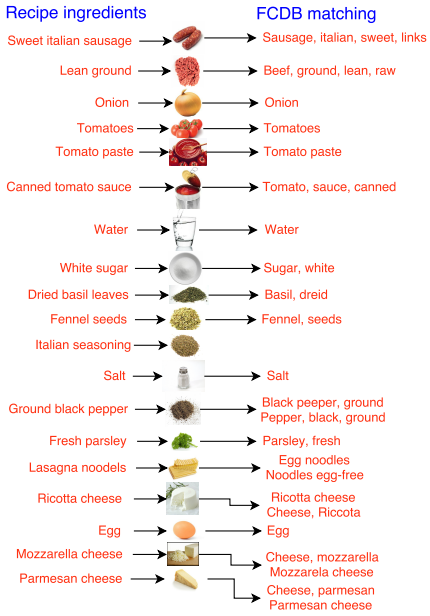


Figure 2: Ingredients matching for "World's best Lasagna".

one ingredient "Italian seasoning". The problem happened because "seasoning" is not annotated here as noun, so we can not continue with the search in the FCDB. We used here the POS tagging which is the part of R programming language. This kind of problem can be solved using some other implementation of POS tagging or some post-processing methods. For other ingredients, we found 18 perfect matches, and for one ingredient, "lasagna noodles", we found most similar match, which is the up close to it, and this happened because "lasagne noodles" is not presented in the FCDB.

We need to mention here that this experiment was carried out without finding the synonyms and mapping the special definite rules.

## 5.2 Experiment 2

Using the 721 lunch recipes, we extracted 1,615 different names of the ingredients that appear in these recipes. In Figure 3, the word cloud of the names of the ingredients found in the recipes is presented. For each ingredient name, using the probability weighted model we found a match in the FCDB that can be in one of the four categories (perfect match, very similar match, similar match, and incorrect match), which we used for evaluation and we manually added to each matching pair. A perfect match is with the same meaning as the ingredient name. A very similar match is the most similar and strongly related to the ingredient name. A similar match is weakly related with the ingredient name. And an incorrect match is incorrect and it can not be used for further analyses. The last two categories appear according to some specific ingredients typical for some cultures and the coverage of the FCDB.

In Figure 4, the pie chart of matching the Internet recipe ingredients with food composition data is presented.

Using the probability weighted model for matching the ingredients, we found 1,210 perfect matches (74.92%), 273 very similar matches (16.90%), 78 similar matches (4.84%) and 54 incorrect matches (3.34%). Let we use the pair ($D_{ingredient}$; $D_{FCDB}$) to describe the match we found, where $D_{ingredient}$ is the ingredient name from the recipe and the $D_{FCDB}$ is the name from the FCDB. For example, some perfect matches are (black olives; olives black), and (fresh ginger; ginger, fresh), very similar matches are (fresh cilantro; spices, coriander seed (cilantro)), and (uncooked egg noodles; egg noodles), similar match is (dry penne pasta; pasta, without egg, dry), and incorrect matches are (angel hair pasta; cake, angelfood,

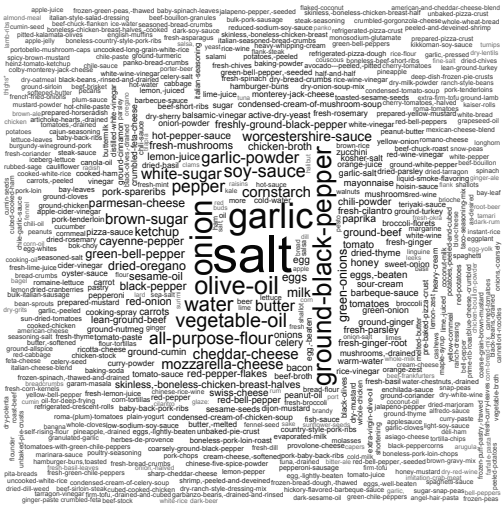


Figure 3: Word cloud of the ingredients.

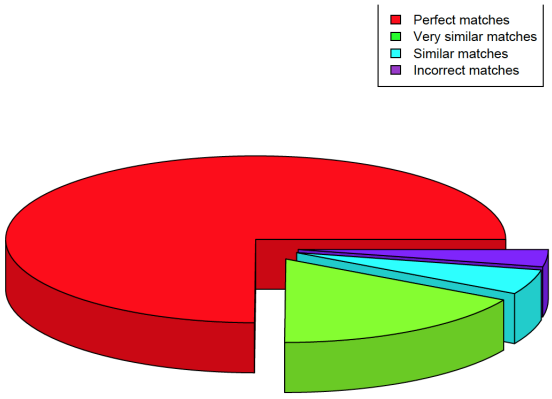


Figure 4: Pie chart of ingredients matching.

commercially prepared), and (dried onion flakes; cereal flakes with dried fruits, type Muesli). The perfect and very similar matches are 91.82% together. They can be used to calculate the nutritional properties on a recipe.

The experiment is done with preprocessed data.

## 6 DISCUSSION

There are some benefits in our method, comparing it with the method that is used to find the most relevant match in (M. Muller et al., 2012). In order to find the most relevant match, they treated the problem as two-class classification problem and to obtain labeled data they asked 6 human assessors to manually evaluate list of ingredients for ambiguous ingredient names. This process ended with 1,515 positively classified instances to which they added the same number again of negatively classified instances. Instead of manually collecting list of ingredients that are positively classified, our method can be used as pre-processing task, and for each of the ingredient can return the relevant ingredient or a list of relevant ingredients, if there are few matching pairs with the maximum weight for the same ingredient. After that, this data can be used for building models, starting with feature selection and then solving two-class or multi-class classification problems. So our method is a benefit to the method proposed in (M. Muller et al., 2012) and can be used as pre-processing step to find the list of ingredients for each ingredient without using the manually evaluation by human assessors. Another benefit is that our method also returned the most similar ingredient that exist in the FCDB and does not require labeled data for supervised learning, the poor choices that appeared are consequence from some ingredients typical for some culture or missing chemical analyses in the FCDB. Also, there are a lot of websites on which we can find recipes by the ingredients we have (MyFridgeFood, 2015; RecipeMatcher, 2015; Supercook, 2015), but using them we can select from a list of ingredients they have, and in the most cases they have only the basic name of the ingredients, without the possibility of using the additional information (the form of the ingredient, or the cooking process). Using them the result is more general, and if we use our method to search the recipe database, the result will contain only the most specific recipes.

We are also working on food image recognition, in order to identify the ingredients in recipes, which is more realistic and challenging task, but the approach is beyond the scope of the paper.

## 7 CONCLUSION

We presented a method, that can be used for matching the recipe ingredients with food composition data. Using this method, we can weight each match between the ingredient name from recipe and food analyses names from FCDBs and then the match with the highest weight is used as the most relevant one. Having this information, we will be able to calculate the nutrition value of each of the recipe which is presented, because for each ingredient used in the recipe we can find the nutritional properties from a FCDBs. Also, this method can be used to weight the ingredients matching, and the weighted data can be used to help more other models, which can be obtained using data mining approaches. This method can be used to explore what is missing in the FCDBs, and this information can be addressed to the chemical laboratories in order to perform food composition data analyses.

We plan to implement this method into a system which will be used for computing the nutritional value of recipes, and to compare the accuracy of the obtained values comparing them with the values from the chemical analyses, which are obtained by chemical analyses of the dishes prepared using the same recipes.

## ACKNOWLEDGEMENTS

This work was supported by the project ISO-FOOD, which received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 621329 (2014-2019).

## REFERENCES

Alani, H., Kim, S., Millard, D. E., Weal, M. J., Hall, W., Lewis, P. H., and Shadbolt, N. R. (2003). Automatic ontology-based knowledge extraction from web documents. *Intelligent Systems, IEEE*, 18(1):14–21.

AllRecipes. Allrecipes website. http://allrecipes.com/. Accessed: 2015-05-04.

A. Voutilainen (2003). Part-of-speech tagging. *The Oxford handbook of computational linguistics*, pages 219–232.

EuroFIR. Eurofir website. http://www.eurofir.org/. Accessed: 2015-05-04.

Han, X., Sun, L., and Zhao, J. (2011). Collective entity linking in web text: a graph-based method. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 765–774. ACM.

H. Greenfield and D. Southgate (2003). *Food composition data: production, management, and use*. Food & Agriculture Org.

J. Freyne and S. Berkovsky (2010). Intelligent food planning: personalized recipe recommendation. In *Proceedings of the 15th international conference on Intelligent user interfaces*, pages 321–324. ACM.

J. Plisson, N. Lavrac, and D. Mladenic (2004). A rule based approach to word lemmatization. *Proceedings of IS-2004*, pages 83–86.

LIRMM. Lirmm. http://data.lirmm.fr/ontologies/food/. Accessed: 2015-05-04.

Mihalcea, R. and Csomai, A. (2007). Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 233–242. ACM.

M. Muller, M. Harvey, D. Elsweiler, and S. Mika (2012). Ingredient matching to determine the nutritional properties of internet-sourced recipes. In *Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2012 6th International Conference on*, pages 73–80. IEEE.

MyFridgeFood. Myfridgefood website. http:// myfridgefood.com/. Accessed: 2015-08-20.

Ontology, B.-F. Bbc - food ontology. http://www.bbc.co.uk/ontologies/fo/. Accessed: 2015-05-04.

RecipeMatcher. Recipematcher website. http:// www.recipematcher.com/. Accessed: 2015-08-20.

R. Real and J. M. Vargas (1996). The probabilistic basis of jaccard's index of similarity. *Systematic biology*, pages 380–385.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the international conference on new methods in language processing*, volume 12, pages 44–49. Citeseer.

Supercook. Supercook website. http:// www.supercook.com/. Accessed: 2015-08-20.

Tian, Y. and Lo, D. (2015). A comparative study on the effectiveness of part-of-speech tagging techniques on bug reports. In *Software Analysis, Evolution and Reengineering (SANER), 2015 IEEE 22nd International Conference on*, pages 570–574. IEEE.

Y.Picó (2012). *Chemical analysis of food: Techniques and applications*. Academic Press.