Extracting Vegetable Information from Recipes to Facilitate Health-Aware Choices

Gijs Geleijnse[1], Thérèse Overbeek[1], Nick van der Veeken[2], and Martijn Willemsen[2]

[1] Philips Research Europe, High Tech Campus 34, 5656 AE Eindhoven, the Netherlands
`firstname.lastname@philips.com`
[2] Eindhoven University of Technology, the Netherlands
`M.C.Willemsen@tue.nl`

1   Introduction

In order to facilitate users of a recipe service to make health-aware meal choices, we focus on the automatic extraction of nutritional information from the text of a recipe. More specifically, the presence of sufficient vegetables in the main meal is addressed in this note. We present an algorithm to extract this information from the recipe texts and explore its use to persuade users to select vegetable-rich meals.

The consumption of a sufficient amount of vegetables is a prominent, universal health recommendation. Moreover, vegetable rich menus are typically low on the number of calories and promote satiety [3]. A 2003 study [2] shows that the average consumption of vegetables in the Netherlands is lagging far behind the generally known governmental recommendation of 200 grams a day.

Obviously, people who are motivated to select a vegetable-rich recipe can inspect the ingredient lists themselves. The mere presence of vegetables in a recipe is fairly easy to recognize, but the quantities of vegetables in terms of grams may be more problematic. For example, when an ingredient list contains entries like *two large onions*, *a can of chopped tomatoes* and *1/2 leek*, it is not straight forward to translate such pieces and units to grams. By making the presence of sufficient vegetables in a meal salient, health-aware choices can be facilitated.

2   A vegetable counting algorithm

We develop an algorithm to automatically identify the number of grams of vegetables in a recipe. With small modifications, the algorithm can be adapted for the American guidelines, where the recommended amount of vegetables is measured in cups or numbers of vegetables and fruits (five a day in UK) [3].
1. *Identify the ingredient list.* The subpart of the recipe containing the listing of ingredients can be detected based on its header or by using a wrapper [1].

---

[3] `http://www.mypyramid.gov/` and `http://www.food.gov.uk/`

ingredient line	product	quantity	unit
2 cans of tomatoes (of 400 grams)	canned tomato	800	grams
2 cloves of garlic	garlic	2	clove
2 cm ginger root	ginger	2	cm
1/2 bag of Radicchio (lettuce, 200 grams)	radicchio	100	grams
2 or 3 small florets of endive (200 g each)	endive	400	g

Table 1. Example of the ingredient lines and the information extracted, translated into English.

2. *Identify the ingredient lines.* Having identified the ingredient list, the list is to be split into ingredient lines. This is done by identifying the separators (e.g. end-of-line markers) that are commonly used in the web site.

3. *Identify the vegetable lines.* Per ingredient line, we extract the longest matching term in a pre-compiled ingredient list. By linking ingredients to food groups, we can isolate the ingredient lines containing vegetables.

4. *Extract the unit and quantity for each vegetable line.* From the ingredient lines that describe a vegetable, we extract the unit and the quantity. The unit is extracted using a list of commonly used unit names.

5. *Estimate the total number of grams of vegetables based on the extracted data.* Having extracted all vegetables, their quantities and units, we use a lookup table to translate the quantity-unit combinations into grams. These units are specified per vegetable. Having computed the weights in grams of all vegetables, we simply determine the sum over all vegetables and divide it by the number of servings.

The described algorithm is evaluated with a set of 3594 recipes from a large Dutch recipe site. Together, these recipes contain 14666 distinct ingredient lines. For 3386 of these lines, the ingredient was identified as a vegetable. To test whether the algorithm is well applicable on these recipes, we analyzed the precision of the output. Out of the 3386 ingredient lines, only two were erroneously marked as containing a vegetable. With respect to the recall, we manually analyzed 100 randomly selected recipes to check for ingredient lines containing vegetables that were not recognized by the algorithm. Within these recipes, five of such vegetable lines were missed. All cases could be resolved by expanding the list of known vegetable names.

Based on this analysis, we can conclude that we can precisely extract the required vegetable information from the text of a recipe. However, an extensive lookup table should be provided to map all packages of vegetables to grams and to estimate the weight of single pieces.

3   A Web Study with Labeled Recipes

We present a web study in which we investigate whether presenting vegetable-rich recipes with a label may persuade users to select such a recipe.

condition	< 200g	> 200g	total	av. grams	st. error
control	55 (52.9%)	49 (47.1%)	104	195.923	3.794
manipulation	59 (56.1%)	46 (43.8%)	105	193.219	3.619
all participants	114 (54.5%)	95 (45.5%)	209	194.571	2.616

Table 2. Number of participants per condition selecting a recipe with < 200 and > 200 grams of vegetables.

Method Two-hundred-and-nine participants (119 females, 19-67 years, mean age = 37, $SD = 11.4$) accepted an e-mail invitation to select a recipe from a list that they would like to prepare *in a couple of days*. We included this clause to exclude practical problems (e.g. plans for today, groceries in stock) and to sketch a scenario that is close to the use of an actual online recipe service.

The 14 recipes presented were selected to be attractive for a broad Dutch audience. The amount of vegetables per serving varied from 150 to 254 grams (average was 199 grams), with half of the dishes containing at least the required amount of vegetables (> 200 grams per serving). The ordering of the recipes was randomized for each participant and dynamically created when accessing the web study. The recipes were compactly represented by a list with the titles and an ingredient list would show up whenever the participant moved the cursor over a recipe title. This allowed us to track the viewing behavior.

There were two experimental conditions. In the manipulation condition, a label resembling the well-known *Choices* logo [4] was added, together with the message *This meal contains enough vegetables*. The message and logo were prominently placed next to the ingredient list. In the control condition, none of the recipes was presented with the message or logo. The assignment of the participants to either condition was alternated, leading to 105 participants in the manipulation condition and 104 participants in the control condition.

In addition, two features of the participants' viewing behavior were derived from the movements of the mouse cursor during the task: the total amount of viewing time per recipe (i.e. opening the ingredient list) and the number of times a recipe was viewed.

Results Analysis shows that participants in both the manipulation condition and the control condition chose both types of recipes about as often, $\chi^2(1) = 0.23, p = .631$. Furthermore, in terms of the average number of grams of vegetables in the recipes selected by the participants, the results again show that the display of the vegetable labels did not increase the choice for vegetable-rich recipes, $t(207) = .516, p = .516$. Thus, the labels did not increase the choice for the healthier recipes when inspecting the entire population.

Viewing behavior was different for the participants who chose a recipe with at least 200 grams of vegetables than for those who chose a recipe with less than 200 grams vegetables. In particular, the participants who chose a recipe with at least 200 grams of vegetables were more frequently, $F(1, 204) = 67.07, p =$

.000, $\eta^2 = .247$, and longer, $F(1, 204) = 31.36, p = .000, \eta^2 = .133$, looking at the healthier recipes than those who chose a recipe with less than 200 grams of vegetables. These findings suggest that those who chose a healthier recipe may have been more intrinsically interested in the healthier recipes and that they may have chosen a recipe more carefully.

4   Conclusions and Future Work

In this paper we have described an algorithm to automatically extract the amount of vegetables in a recipe. This information can be used to enrich a recipe website as a means to persuade its users to select healthier meals.

We have shown that the vegetable counting algorithm gives precise results and have applied the algorithm to label the recipes that contain the recommended amount of vegetables. This label can facilitate the choice of vegetable-rich meals, as this information cannot directly be derived from the recipe text. However, the web study using labeled recipes showed that the label did not lead to more vegetable-rich recipe selections overall. The choice for a particular recipe seemed to be driven more by the participants liking for a recipe than by whether or not it was healthy. The study was not designed to measure the prolonged use of the logos. Familiarity with the nutritional support system and its use may have an effect on the users' browsing and selection behavior.

For future work, an at-home study where the prolonged use of a nutritional support system with vegetable labeling can give additional insights on meal selections over a period of time. The use of the system with labeling may improve the awareness of one's eating behavior over time. The vegetable information of the selected recipe can also provide the user with feedback on his consumption patterns. Another route to explore is the inclusion of vegetable information in a recipe recommender system. By combining user preferences, eating behavior and health information, we can present the user with attractive, health-aware recipe suggestions. Such recommendations may also improve a user's vegetable intake.

References

1. V. Crescenzi and G. Mecca. Automatic information extraction from large websites. *Journal of the ACM*, 51(5):731 – 779, 2004.
2. K. Hulshof, M. Ocke, C. v. Rossum, E. Buurma-Rethans, H. Brants, J. Drijvers, and D. t. Doest. Results of the national food consumption survey 2003, 2004.
3. B. Rolls, J. Ello Martin, and B. Tohill. What can intervention studies tell us about the relationship between fruit and vegetable consumption and weight management? *Nutrition Reviews*, 62:1 – 17, 2004.
4. E. L. Vyth, I. H. M. Steenhuis, S. F. Mallant, Z. L. Mol, J. Brug, M. Temminghoff, G. I. Feunekes, L. Jansen, H. Verhagen, and J. C. Seidell. A front-of-pack nutrition logo: A quantitative and qualitative process evaluation in the netherlands. *Journal of Health Communication*, 14(7):631 – 645, 2009.