Clustering for Closely Similar Recipes to Extract Spam Recipes in User-generated Recipe Sites

Shunsuke Hanai
Konan University
Okamoto8-9-1
Higashinada-ku Kobe, Japan
m1424007@center.konan-u.ac.jp

Hidetsugu Nanba
Hiroshima City University
Ozukahigashi 3-4-1
Hiroshima 731-3194 Japan
nanba@hiroshima-cu.ac.jp

Akiyo Nadamoto
Konan University
Okamoto8-9-1
Higashinada-ku Kobe, Japan
nadamoto@konan-u.ac.jp

ABSTRACT

Nowadays, many user-generated recipe sites are accessible on the internet. On user-generated recipe sites, however, are various spam recipe pages that describe closely similar recipes requiring special cooking equipment, with no preparation explanations. These spam recipes are not useful for users. In fact, they impede user's recipe searches. In this paper, we target closely similar recipes as a first step in extracting spam recipes. If user search results could be classified to identify closely similar recipes, user's recipe searches would be easier and more productive. Clustering tools of many kinds are proposed, but it is difficult to cluster closely similar recipes using only existing clustering tools because recipe sites have a unique page structure comprising a title, ingredients, directions (preparation instructions), and comments. The importance of words from each part differs. We propose a clustering method for user-generated recipe sites based on page structure and important words. Next, we conducted an experiment to measure the benefits of our proposed method. The result of experiment presents the benefits of our proposed method which classify the closely similar recipes.

Categories and Subject Descriptors

H.4.m [Information Systems]: Miscellaneous; I.7.m [Document and Text Processing]: Miscellaneous

General Terms

Algorithms

Keywords

User-generated recipe, Clustering, Closely similar recipes

1. INTRODUCTION

Many user-generated recipe sites are accessible on the internet today, such as Food.com[1](U.S.), Mis Recetas[2](Hispanic), Beitai Chufang[3](China) and Cookpad[4](Japan). These recipe sites have many users who post their recipes freely, providing recipes of many kinds for themes such as home cooking, party cooking, and diet foods. Many users also search for recipes for their meals from the recipe sites. For example, the number of monthly users of Cookpad is approximately 44,930,000 (July, 2014). The number of recipe pages is about 2,120,000 (August 2015).

When a user searches for a recipe, the user poses queries of two types, typically incorporating food names such as beef stew or lasagna, and ingredient names such as chicken, cabbage, or onions. Maruha-Nichiro Holdings, a famous food company in Japan, investigated which query is better used when users seek recipes posted on recipe sites [8]. In fact, ingredient keywords are used more often than food name keywords. Ingredient names account for 75% of all keywords. Results show that users who search for recipes before deciding a menu account for about 57% of all users. Users who search for a recipe after they have chosen a menu are about 26% of all users. Therefore, when users use recipe sites, they input an ingredient name as a query. However, when users search for a recipe using an ingredient name, they are often deluged by inconvenient recipes of many kinds that are closely similar recipes, but which use special cooking equipment, and which include no preparation explanation whatsoever. These inconvenient recipes are not useful for users. Actually, they impede user's recipe searches. These closely similar recipes obstruct user searches because of their sheer number on user-generated recipe sites. They are, in fact, spam recipes. We regard extraction of these spam recipes as an important and valuable service. Especially, deliberately or accidentally, numerous closely similar recipes are posted among the user-generated recipes. For example, when a user inputs “Tomato and Chicken” in Cookpad, a famous Japanese user-generated recipe site, the results extend to more than 10,000 pages. Numerous closely similar pages are found. That “information overload" confuses users.

Moreover, a user who searches for a recipe usually does not select a high-ranking recipe from the search results, re-

[1] Food.com http://www.food.com/
[2] Mis Recetas http://www.misrecetas.com/
[3] Beitai chufang http://www.beitaichufang.com/
[4] Cookpad http://cookpad.com/

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

iiWAS2015, 11-13 December, 2015, Brussels, Belgium.

Copyright 2015 ACM 978-1-4503-3491-4/15/12...$15.00..

acting better than one might with usual web searches [10]. Data show that users typically compare multiple recipes. Given many similar recipes included in the search results, it would be difficult to compare multiple recipes. When users compare similar recipes, they must better understand the different points of similar recipes. That need for comparison imposes a great burden on users. A system classifying the results of user searches according to similar pages in real time would be beneficial for users. As described herein, we propose a method that classifies the user search results according to closely similar pages based on the page structure and the types of important words. Clustering tools of many kinds have been proposed, but it is difficult to classify closely similar recipes using existing clustering tools because recipe sites have a unique page structure that includes the title of a dish, with ingredients, directions (preparation instructions), and comments. These parts of pages differ in their roles, importance, and meaning. As described in this paper, we propose a means of classifying the results of a user search according to closely similar pages. Next, we conducted an experiment to assess the benefits of our proposed method.

The remainder of this paper is organized as explained below. First, Section 2 presents a summary of related works. Section 3 presents a method for clustering the spam recipes. Section 4 presents discussion related to experimental evaluation, along with evaluation results. Finally, Section 5 presents conclusions and expectations for future work.

2. RELATED WORK

Along with the growth of recipe-sharing services, many studies have examined recipe recommendation processes. Wagner et al. [14] proposed a system that tracks a user's cooking activities with sensors in kitchen utensils and which recommends healthy recipes that might increase the user's cooking competence. Svensson et al. [11] applied their idea of social navigation for recipe recommendation. Specifically, they assigned users to groups based on their explicit preference information such as ingredients, fat level, and time to cook. Geleijnse et al. [2] designed a prototype of a personalized recipe recommendation system, which suggests recipes to users based on their past food selections and nutrition intake. Lawrence et al. [7] proposed a product recommendation system for grocery shopping using collaborative filtering to rate products based on the user's prior purchase behavior. Harvey et al. [5] explained factors of selecting recipes by assessing the results of long-term research to recommend recipes that match a user's preferences. However, we particularly examine recipe similarity, not personalization. We aimed to be helpful for user's search recipes by extracting spam recipes.

Teng et al. [12] proposed ingredient recommendation systems using ingredient networks. They constructed networks of two types to capture the relations among ingredients: ingredient complements and substitute ingredients. Pinxteren et al. [13] identified important features and extracted them from the recipe texts. They calculated the degree of similarity among recipes, and changed to healthy recipes. Shidochi et al. [9] proposed a method to identify replaceable ingredients by matching the cooking actions that correspond to ingredient names from recipe texts. Forbes et al.[1] apply a matrix factorization method to recipe recommendation. Experimental results demonstrated that the algorithm not only improves the recommendation accuracy; it is also useful for swapping ingredients and creating new recipes. These studies have focused on ingredients. However, these studies specifically examined substitute ingredients and substitute recipe recommendation systems: the extraction of similar recipes is not the goal.

3. CLOSELY SIMILAR RECIPE CLUSTERING

From our previous study[3][4], we revealed the following four points for the clustering of closely similar recipes.

- No image is necessary to judge recipe similarity.
- Food names, ingredients, seasonings, and cooking methods in a title are important words. Especially, users care about food names and cooking methods in the title when judging closely similar recipes.
- Main ingredients and the main seasoning in an ingredient list are important.
- Sizzle words in a title are not necessary to judge the recipe similarity.

We propose clustering methods based on four points to extract closely similar recipes. The flow of clustering for closely similar recipes is the following:

1. The system searches the recipe pages using multiple ingredients from the user input.
2. It extracts the title and ingredient list from search results.
3. It extracts nouns from the results of (2) and performs classification by food name, cooking method, ingredient, and seasoning, depending on our food database.
4. It classifies recipe pages according to the kind of food based on the food name and cooking method in title. Here, we use Repeated Bisection method[15] as a clustering method.
5. It calculates feature values of ingredients and seasonings in titles and ingredient lists.
6. It classifies results of (4) according to closely similar recipes using feature values of ingredients and seasonings in the titles and ingredient lists. At this time, we use words without user input keywords because input keywords certainly include all recipe pages. The keywords are not regarded as characteristic words. If the keywords include clustering-based words, then the keywords might become central words of a cluster. We can not extract closely similar recipes. Therefore, we delete the keywords when classifying the recipes.

3.1 First Clustering Based on Food Names and Cooking Methods in the Title

Our previous study demonstrated that the food name and cooking method in the title is the most important element to judge closely similar recipe. We first classify recipe pages based on food names and cooking methods in the title.

The target data are search results of recipe pages using user input multiple ingredient such as “pork and onion,” and “chicken and tomato.” We extract titles. Next, we extract food names and cooking methods using our food database, which includes food names and cooking methods. When we classify recipe pages, we use Repeated Bisection [15] which is a method used for bayon [5] and CLUTO [6]; it is a kind of K-means method. After the first clustering, the recipe pages are clustered by food names or cooking methods included in the title area.

3.2 Second Clustering Based on Ingredients and Seasoning

The first clustering uses only the food name and cooking method. There are many different meals but they are in the same cluster. For example, one cluster is related to curry, which is a food name. The elements of the cluster include “Vegetable curry,” “Tomato curry,” “Pork curry,” “Soft pork curry,” “Chicken curry,” and “Chicken spicy curry.” They are not closely similar recipes. Next, we classify them again based on ingredients and cooking methods.

Our previous study demonstrates that words in a title are more important than other words in a passage. Calculating the weight of words is necessary based on their position of appearance on a page. Furthermore, a surprising degree of ingredients in a food name is important because unusual ingredients for the food should be regarded as recipe characteristics. For example, for the recipe title “Lasagna,” some ingredients are tomato, pork, onion, cheese, and tofu. In this case, the tofu is an unusual lasagna ingredient. Therefore, that unusual ingredient becomes a recipe characteristic. Then we calculate the surprising degree of ingredients using our surprising degree S-RF-IIF which is based on the Recipe Frequency-Inverted Ingredient Frequency (RF_IIF) [6]. S_RF_IIF calculates a value reflecting the degree of surprise which might be associated with ingredient inclusion based on the term frequency and the term-appearance position. Expression of $S\_RF\_IIF_{i,m}$ as follows:

$$S\_RF\_IIF_{i,m} = \alpha \log \frac{|R_m|}{|R_{i,t,m}|} + \beta \log \frac{|R_m|}{|R_{i,o,m}|} \qquad (1)$$

The given i is an ingredient name and m is a food name. In addition, $|R_m|$ denotes the number of recipes of food m, $|R_{i,t,m}|$ denotes the number of recipes of food m, which includes ingredient i in title of food m. $|R_{i,o,m}|$ denotes the number of recipes of food m, which includes ingredient i in passages except title of recipe page of food m. In that equation, α signifies a weight that is dependent on a term's appearance position in title, and β signifies a weight that is dependent on a term's appearance position in ingredient list. In this paper, we regard $\alpha = 1.0$ and $\beta = 0.5$ by experiment of parameter determination.

Then we classify the results of the first clustering using Repeated Bisection based on ingredients and seasonings, which are feature values calculated using S-RF-IIF.

4. EXPERIMENTS

Table 1: Condition of Experiment 2.

Query	Number of recipe pages	Number of cluster in clustering
Pork ∩ Egg Plant	5,885	135
Pork ∩ Onion	28,525	230
Pork ∩ Radish	8,446	146
Chicken ∩ Potato	9,147	142
Avocado ∩ Tomato	5,284	98

We measure the accuracy of our proposed method. The datasets are shown in Table 1. We discuss the results divided into the first clustering and second clustering.

4.1 First Clustering Based on Food Name and Cooking Method

First, we classify recipes based on the food name and cooking method. Table 2 presents examples of the results for each of the top three clusters. The results demonstrated that we can find the clusters are divided into the food name or cooking method. However, we can find that a cluster divided into food or cooking methods includes various main ingredients or seasonings. For example, the query of “Pork and Onion” in the first cluster includes vegetable curry, tuna curry, and pork curry. In this case, the cluster is divided into the food name “curry,” but the main ingredients differ; they are not closely similar recipes. Therefore, we must classify smaller clusters using ingredients and seasonings.

4.2 Second Clustering Based on Ingredient and Seasoning

Next, we classify the results of the first clustering based on ingredients and seasonings by considering the appearance points. Table 3 presents examples of the results of the top three clusters of each cluster 1 in Table 2. The results demonstrate that the same food names with different main ingredients are classified into different clusters. For example, in the first clustering at query for “Pork and Onion”, the different main ingredient in the same cluster, but as a result of the second clustering, these recipes are classified into different clusters. Almost all closely similar recipes are classified into the same cluster. We can then classify the closely similar recipes using our proposed method. We discuss the precision of the results. The results of precision are dispersion (see Figure 1). In Figure 1, the horizontal axis presents the values of precision. The vertical axis shows the number of clusters. The value of precision having the greatest number of clusters is 0.6 in the five graphs. However, the value of precision is 0. It has a large number of clusters for two reasons: (1) general seasonings, which are water and salt, become the center words of the cluster; (2) general ingredients of a dish, such as carrots and potatoes in curry, become center words. We should consider the two problems explained above.

5. CONCLUSION

This paper proposed a clustering method to extract closely similar recipes in user generated recipe sites such as Food.com, Mis Recetas, Beitai Chufang, and Cookpad. Our proposed method is twice clustering, with specific examination of the page structure and types of important words. The first clustering classifies user search results based on the food name

[5] Bayon https://code.google.com/p/bayon/wiki/Tutorial_en
[6] CLUTO http://glaros.dtc.umn.edu/gkhome/cluto/cluto /overview

Table 2: Results of First Clustering

Query : Pork AND Eggplant		
Cluster1, Dish name: Stir - fry	Cluster2, Dish name: Saute	Cluster3, Dish name: Pasta
Recipe title	Recipe title	Recipe title
Eggplant miso stir-fry	Pork ginger saute	Tomato pasta with eggplant and ground meat
Eggplant stir-fry with miso	Pork saute with mustard flavor	Eggplant meat pasta
Meat and eggplant stir-fry with oyster sauce	Simply! Pork saute with special sauce	Pasta with bitter melon, eggplant, and ground meat
Chopped pork and eggplant stir-fry of thick and delicious	Pork saute!! Balsamic sauce	Tomato pasta with eggplant and ground meat
Simply! Pork belly stir-fry with gochujang	Pork saute with spinach cream sauce	Tomato cream pasta with milk
Query : Pork AND Onion		
Cluster1, Dish name: Curry	Cluster2, Dish name: Roll	Cluster3, Dish name: Meatloaf
Recipe title	Recipe title	Recipe title
Curry with plenty of summer vegetables	Plenty of veggies! Basic spring roll	Healthy! Meatloaf
Japanese-style tuna curry	Spring roll at home	Made with ground meat! Meatloaf
Curry flavor of pork and apple	Basic spring roll!!	For a party! Juicy meat loaf
Low price! Pork belly and cartilage curry	Made at home! Home-made spring roll	Made in the microwave!! Meatloaf
Pork curry of Larger vegetables	Mini spring roll	Cheese meat loaf with celery and raisins
Query : Pork AND Radish		
Cluster1, Dish name: Simmered	Cluster2, Dish name: Marinade	Cluster3, Dish name: Sukiyaki
Recipe title	Recipe title	Recipe title
Simmered radish and ground meat	Marinated Pork &Veggies for a feast	Pork kimchi sukiyaki
Simmered pork	Refreshing taste! Marinated pork	Not a beef! Pork sukiyaki
Chinese-style simmered pork and root vegetables	Marinated thinly sliced pork meat	For Beauty! Healthy sukiyaki bowl
Simmered radish in a pressure cooker	Pork marinated in pickled plum & miso	Pork sukiyaki with radish and egg
Lightly simmered radish	Easy Japanese-style marinated pork & radish	Pork sukiyaki part2! Easy!
Query : Chicken AND Potato		
Cluster1, Dish name: Stew	Cluster2, Dish name: Risotto	Cluster3, Dish name: Soup
Recipe title	Recipe title	Recipe title
Tomato stew with chicken meatball	Tomato risotto	Potato potage soup
Tomato stew with rich flavor and sweet chicken	Very easy!! Baked cheese curry risotto	Easy potato and enoki mushroom potage soup
Chicken and vegetable stew	Chicken thigh and seasonal vegetables risotto	Burdock potage soup
Cream stew with soy milk	Potato risotto	Plenty of veggies! Potage soup
Cream stew	Spring!! New potatoes and chicken risotto	Spring cabbage potage soup
Query : Avocado AND Tomato		
Cluster1, Dish name: Salad	Cluster2, Dish name: Sandwich	Cluster3, Dish name: Omelette
Recipe title	Recipe title	Recipe title
Tuna and avocado salad	Avocado sandwich	Avocado omelette
Tuna and avocado salad	Avocado and egg sandwich	Omelette(avocado & tomato)
Basil flavor salad with avocado and mozzarella	Avocado and Tomato! Veggie sandwich	Fluffy omelette with avocado and chicken
Avocado and tuna salad	For diet! Avocado sandwich	Italian-style omelette with avocado
Japanese style salad of avocado and cream cheese	Sandwich with fresh vegetables	Healthy! Veggie omelette

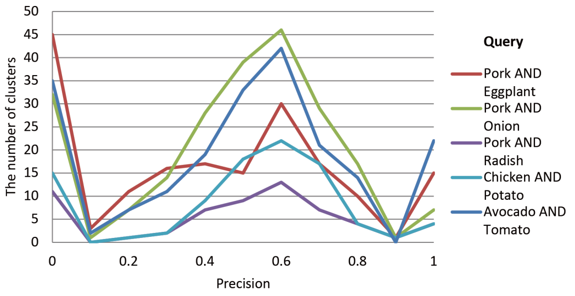

Figure 1: Dispersion of Precision

and cooking method in the title. The next clustering classifies the results of the first clustering based on ingredients and seasonings. When calculating the second clustering, we use the ingredient weight based on appearance points. Our experiment shows that our proposed method can find clusters of closer similar recipes.

In the near future, we expect to consider the cooking flow in addition to the values of ingredients and seasonings to cluster similar recipes. We expect to produce a user interface to browse with similarity clustering.

6. ACKNOWLEDGMENTS

This work was partially supported by JSPS KAKENHI grant number 26330347. In this paper, we used recipe data provided by Cookpad and the National Institute of Informatics.

7. REFERENCES

[1] P. Forbes and M. Zhu. Content-boosted matrix factorization for recommender systems: experiments with recipe recommendation. In Proceedings of the fifth ACM conference on Recommender systems, pages 261–264. ACM, 2011.

[2] G. Geleijnse, P. Nachtigall, P. van Kaam, and L. Wijgergangs. A personalized recipe advice system to promote healthful choices. In Proceedings of the 16th international conference on Intelligent user interfaces, pages 437–438. ACM, 2011.

[3] S. Hanai and A. Nadamoto. Clustering for similar recipes by using cooking ingredient. In Proceedings of the IEICE conference on Data Engineering and Food Media, pages 47–52, 2014. (in Japanese).

[4] S. Hanai, H. Nanba, and A. Nadamoto. Clustering for closely similar recipes to extract spam recipe. In IPSJ SIG Technical Reports, pages 1–7, 2014. (in Japanese).

[5] M. Harvey, B. Ludwig, and D. Elsweiler. You are what

Table 3: Result of Second Clustering

Query : Pork AND Eggplant in cluster1 of table2

Cluster1	Cluster2	Cluster3
Recipe title	Recipe title	Recipe title
Pork & eggplant stir-fry with gochujang	Eggplant and pork stir-fried in oyster sauce	Miso stir-fried with pork & eggplant & green pepper
Thickly stir-fry with eggplant and pork	Eggplant and pork belly stir-fried in oyster sauce	For lunch! Miso stir-fried with eggplant
Gochujang stir-fry with eggplant and pork	Oyster sauce stir-fried with eggplant, green pepper	Miso stir-fried with plenty of vegetables
Rich stir-fry with pork and eggplant	Eggplant and pork belly stir-fried in miso oyster sauce	Pork miso stir-fried with eggplant and green pepper
Simply! Pork belly stir-fry with gochujang	Sweet miso stir-fried with Eggplant and green pepper	Miso stir-fries with pork, eggplant and green pepper

Query : Pork AND Onion in cluster1 of tabel2

Cluster1	Cluster2	Cluster3
Recipe title	Recipe title	Recipe title
Thickly! Pork cartilage curry	Delicious! Pork curry	Beat the summer heat! summer vegetables curry
Low price! Pork belly and cartilage curry	In a pressure cooker! Pork curry	Plenty of summer vegetables curry
Delicious home made! Pork cartilage curry	Very easy! Pork curry in 15 minutes	Keema curry with summer vegetables
Delicious curry with cartilage and lotus root	Pork beans curry	Summer vegetables curry!!
Collagen rich pork cartilage curry	Pork ginger curry	Plenty of summer vegetables! Tomato curry

Query : Pork AND Radish in cluster1 of table2

Cluster1	Cluster2	Cluster3
Recipe title	Recipe title	Recipe title
Addictive chewy! Simmered pig cartilage and radish	Simmered pork belly, radish and burdock	Easy pressure cooker recipe! Simmered ground pork
Simmered pig cartilage and radish	Simmered pork belly and root vegetables	Simmered radish and ground pork
Simmered radish and pig cartilage	In a rice cooker! Simmered pork and burdock	Simmered Chinese cabbage and pork meat balls
Melting! Simmered pig cartilage and vegetables	Delicious! Simmered pork and radish	Simmered radish and ground pork
Simmered pig cartilage	Easy! Simmered pork belly and root vegetables	Simmered burdock, tofu and ground pork

Query : Chicken AND Potato in cluster1 of table2

Cluster1	Cluster2	Cluster3
Recipe title	Recipe title	Recipe title
Chicken soy milk stew	Warm creamy stew	Chicken tomato stew
Stew using soy milk	Cream stew with plenty of autumn vegetables	Rich taste! Easy warm tomato stew
Cream stew with plenty of veggies and soy milk	Simple! Cream stew for adult	Warm tomato stew
Soy milk cream stew	Mellow chicken cream stew	Tomato stew with plenty of autumn vegetables
Soy milk stew	Cream stew in a pressure cooker	Tomato stew

Query : Avocado AND Tomato in cluster1 of table2

Cluster1	Cluster2	Cluster3
Recipe title	Recipe title	Recipe title
Tuna and avocado salad	Avocado and cottage cheese salad	Easy! Salmon and avocado salad
Tuna and avocado salad	Easy, avocado and cottage cheese salad	Beautiful skin salad with avocado and salmon
Tuna, tomato and avocado salad	Avocado and cottage cheese salad!	Health & Beauty! Salmon and avocado salad
Plenty of vegetables! Tuna and avocado salad	Avocado, tomato and cottage cheese salad	Tomato, salmon and avocado salad
Tuna & avocado salad with wasabi soy sauce	Japanese-style salad with avocado, tomato and cheese	Refreshing salad with avocado and salmon

you eat: Learning user tastes for rating prediction. In Proceedings of String Processing and Information Retrieval, pages 153–164. Springer, 2013.

[6] K. Ikejiri, Y. Sei, H. Nakagawa, Y. Tahara, and A. Ohsuga. A proposal of calculation method of surprising value of recipe based on ingredient. Proceedings of IEICE technical report. Data engineering, 113(214):1–6, 2013. (in Japanese).

[7] R. D. Lawrence, G. S. Almasi, V. Kotlyar, M. Viveros, and S. S. Duri. Personalization of supermarket product recommendations. the Journal of Data Min. Knowl. Discov., pages 11–32, 2001.

[8] Maruha-Nichiro. Investigation of cooking recipes, 2013. http://www.maruha-nichiro.co.jp/news_center/research/pdf/20130227_recipe_cyousa.pdf [Online; accessed 11-August-2015](in Japanese).

[9] Y. Shidochi, T. Takahashi, I. Ide, and H. Murase. Finding replaceable materials in cooking recipe texts considering characteristic cooking actions. In Proceedings of the ACM Multimedia 2009 Workshop on Multimedia for Cooking and Eating Activities, pages 9–14. ACM, 2009.

[10] Y. Sugiyama, Y. Yamakata, and K. Tanaka. Summary of similar recipe for the recipe data as a procedure information and discovery of important differences. In Proceedings of DEIM Forum 2013 D3, volume 5, 2013. (in Japanese).

[11] M. Svensson, K. Höök, J. Laaksolahti, and A. Waern. Social navigation of food recipes. In Proceedings of the SIGCHI conference on Human factors in computing systems, pages 341–348. ACM, 2001.

[12] C.-Y. Teng, Y.-R. Lin, and L. A. Adamic. Recipe recommendation using ingredient networks. In Proceedings of the 4th Annual ACM Web Science Conference, pages 298–307. ACM, 2012.

[13] Y. van Pinxteren, G. Geleijnse, and P. Kamsteeg. Deriving a recipe similarity measure for recommending healthful meals. In Proceedings of the 16th international conference on Intelligent user interfaces, pages 105–114. ACM, 2011.

[14] J. Wagner, G. Geleijnse, and A. van Halteren. Guidance and support for healthy food preparation in an augmented kitchen. In Proceedings of the 2011 Workshop on Context-awareness in Retrieval and Recommendation, pages 47–50, 2011.

[15] Y. Zhao and G. Karypis. Comparison of agglomerative and partitional document clustering algorithms. Technical Report 02-014, University of Minnesota, 2002.