Extracting Food Substitutes From Food Diary via Distributional Similarity

Palakorn Achananuparp
Singapore Management University
Singapore
palakorna@smu.edu.sg

Ingmar Weber
Qatar Computing Research Institute
Qatar
iweber@qf.org.qa

ABSTRACT

In this paper, we explore the problem of identifying substitute relationship between food pairs from real-world food consumption data as the first step towards the healthier food recommendation. Our method is inspired by the distributional hypothesis in linguistics. Specifically, we assume that foods that are consumed in similar contexts are more likely to be similar dietarily. For example, a turkey sandwich can be considered a suitable substitute for a chicken sandwich if both tend to be consumed with french fries and salad. To evaluate our method, we constructed a real-world food consumption dataset from MyFitnessPal's public food diary entries and obtained ground-truth human judgements of food substitutes from a crowdsourcing service. The experiment results suggest the effectiveness of the method in identifying suitable substitutes.

CCS Concepts

•Information systems → Recommender systems;

Keywords

Food recommendation; food substitutes; distributional similarity; food diary; food journal; MyFitnessPal

1. INTRODUCTION

Forming and maintaining healthy eating habits is important to individuals' long-term physical well-being. However, despite the availability of numerous dietary guidelines, few people are able to do so as demonstrated by the prevalence of chronic diseases such as obesity and type-2 diabetes. Arguably, part of the reason for such failure is that these guidelines are one-size-fits-all suggestions, making them difficult to be adopted habitually by individuals. In contrast to dietary guidelines, suggestions tailored to specific individuals from recommender systems may be more effective at facilitating incremental behavior change. More specifically, by learning about users' dietary behavior through data from mobile food consumption tracking apps such as MyFitnessPal (MFP), the systems can nudge the users towards "similar but healthier" alternatives by recommending food substitutes personalized to the users' current dietary needs and preferences. In this work, we explored a data-driven approach to extracting food substitutes from personal food consumption data as the first step into the healthier food recommendation. Thanks to the rise of self-monitoring practices enabled by mobile and wearable technology, we turned to a wealth of public food consumption data created by MFP users.

One	Calories	Protein	Fat	Carbs	Fiber	Sodium
Compliments - Mini Brownies, 40 grams (2 brownies)	179	2	9	21	1	100
Grimm's - Lean Turkey Pepperoni, 63 grams (1 stick)	105	13	6	2	0	609
Compliments - Paradise Trio Frozen Fruit, 154 grams	66	1	0	19	2	0
Fruit - Strawberries, 138 grams	44	1	0	11	3	1
Cheetos - Crunchy Cheetos, 16 grams (37 pieces)	90	1	6	10	0	141
Heritage - Hillbilly Kettle Corn, 9 grams	43	1	2	5	1	106
Quick Tools	527	19	23	68	7	957
Four						
Fruit - Grapes, 108 grams	72	1	0	18	1	2
Fruit - Oranges, General, 146 grams	72	1	0	18	3	0
Heritage - Hillbilly Kettle Corn, 9 grams	43	1	2	5	1	106
Quick Tools	187	3	2	41	5	108
Totals	2,015	129	59	246	29	2,459
[redacted] Daily Goal	2,020	126	65	233	25	2,300
Remaining	5	-3	6	-13	-4	-159
	Calories	Protein	Fat	Carbs	Fiber	Sodium

Figure 1: Screenshot of a food diary on MFP.

MyFitnessPal (MFP) is a popular mobile app and website for fitness and health with 80 million registered users[1]. One of its core features is an online food diary which helps users track their food consumption to achieve specific health goals, such as losing, gaining, or maintaining weight. Each food diary page consists of a sequence of meals where each meal contains a collection of food entries and nutrition information. As we can see from the sample diary page shown

HealthRecSys 2016 Boston, Massachusetts USA

ACM ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

[1] https://en.wikipedia.org/wiki/MyFitnessPal

in Figure 1, the user has logged 9 food entries across 2 meals (named "One" and "Four"). When logging a new entry into a diary, users can either enter a new food entry and nutritional values or search for existing food entries shared by other users in the food database. Users can control the diary sharing setting such that their diaries can be viewed by anyone ("Public"), their friends ("Friends Only"), or only the users themselves ("Private").

Our main assumption of the substitution relationship between foods is inspired by the distributional hypothesis in linguistics: words that occur in the same contexts tend to convey similar meanings. By applying the same notion to food consumption, we hypothesize that foods consumed in similar contexts are more likely to be a substitute of each other. More specifically, our method is based on the vector space models of semantics [10] commonly used in related natural language processing (NLP) tasks, such as word similarity and analogy recovery.

Past studies have investigated the applications of recommender systems in food and cooking domains. One of the most common tasks explored by researchers is cooking recipe recommendation [4, 6, 5] where popular recommendation algorithms, such as collaborative filtering [4] and matrix factorization [6, 5] were employed to predict ratings of cooking recipes. Others have focused on extracting substitutable ingredients from recipes using network analysis [9] and statistical approaches [1]. While many past studies relied on recipe ratings data collected through recipe-sharing websites such as AllRecipes.com, to the best of our knowledge, our work is the first to study the food substitute extraction problem using a real-world self-reported food consumption data. In addition, we identified the substitution relationship directly from the food consumption data instead of relying on external knowledge sources[9, 1]. Lastly, we evaluated the effectiveness of the method on the human-labeled dataset of food substitutes constructed through an online crowdsourcing service.

The rest of the paper is organized as followed. First, we describe our method in Section 2. In Section 3, we describe the procedures to collect and process data. Then, we present the experimental evaluation in Section 4 and discuss the results in Section 5. Lastly, we conclude the paper in Section 6.

2. FOOD SUBSTITUTE EXTRACTION

Our approach to food substitute extraction is based on the assumption that foods consumed in similar contexts tend to be similar dietarily, i.e., they are a suitable substitute of each other. In this work, the contexts comprise other foods consumed together in the same meals. Specifically, the substitutability between two foods is measured by the cosine similarity or dot product between their vector representations. We explore two explicit representation methods commonly used in similar NLP tasks [8]: Food-context matrix (PPMI matrix) and Singular Value Decomposition (SVD).

2.1 Food-Context Matrix (PPMI Matrix)

First, given the food consumption data, we constructed a food-context matrix M where each row represents a food item $f \in V_f$ and each column represents a context $c \in V_c$, where V_f and V_c are the sets of observed food items and contexts, respectively. Each cell M_{ij} represents the association between the food item f_i and the context c_j indicated by Positive Pointwise Mutual Information (PPMI) $ppmi_{ij}$ in Equation 1. PPMI has been shown to perform better than other weighting approaches in semantic similarity tasks [8].

$$ppmi_{ij} = \max(\log \frac{\#(f_i, c_j) * |D|}{\#(f_i) * \#(c_j)} * \sqrt{\max(\#f_i, \#c_j)}, 0) \quad (1)$$

where D denotes the set of all food-context pairs, $\#(f_i, c_j)$ denotes the number of times the pair (f_i, c_j) occurred in D, and $\#(f_i)$ and $\#(c_j)$ are the number of times f_i and c_j occurred in D, respectively. To obtain D, we define the contexts of food item f_i as the food items that are consumed in the same meal as f_i. As PMI is known to bias towards infrequent events, we mitigate this problem by adopting a variant of PMI_{sig} proposed in [2].

Finally, we measure the similarity between two food items by computing the cosine similarity between the corresponding row vectors in the food-context matrix M. Food pairs with higher cosine similarity are more likely to be a suitable substitute of each other than those with lower cosine similarity.

2.2 Singular Value Decomoposition (SVD)

One drawback of PPMI matrix is the sparsity of the food-context matrix M which affects the performance of similarity measurements. One common way to improve the similarity computations is to perform dimensionality reduction through truncated Singular Value Decomposition (SVD) as proposed in Latent Semantic Analysis (LSA) [3]. Basically, SVD decomposes the matrix M into the product of three matrices $U \Sigma V^T$, where U and V are orthogonal and Σ is a diagonal matrix of singular values. Let Σ_k be the diagonal matrix formed by selecting only the top k singular values and U_k and V_k be the matrices formed by selecting the corresponding columns from U and V, so that $M_k = U_k \Sigma_k V_k^T$ is a rank-k approximation of M. Then, the similarity between two food items is measured from the dot product between two corresponding row vectors of M_k.

3. DATA COLLECTION & PROCESSING

In this section, we describe the procedures to collect food consumption data from MFP food diary pages, preprocess food entries in the diaries, and construct the food-context matrix and its low-dimensional representation.

3.1 Obtaining Food Diaries

We collected food diary data by web scraping MFP public food diary pages. First, we identified approximately 100,000 seed users who are a member of at least one of the 10 most popular groups in MFP communities. For each user whose food diary pages were publicly viewable, we retrieved up to the last 180 days of their food diaries (until March 2015 when the data collection took place). In total, 587,187 food diary pages of 9,896 users were retrieved. On average, each user has logged 59.3 days of diaries (S.D. = 54.6, median = 42) or 652.9 food entries in total (S.D. = 774, median = 366). The average age of users in the dataset is 35.6 years old (S.D. = 10.17). The vast majority of users are female (82%) who live in the United States. The gender distribution of our dataset is similar to that of a larger sample used in [7].

3.2 Preprocessing Food Entries

After retrieving the food diaries, we parsed the HTML content of the diary pages to extract individual meals and

# users	# meals	# unique food entries
9,896	1,919,024	71,7175

Table 1: MFP Dataset

food entries. Since food entries are described in free text, different entries may refer to the same dish. For example, both "toasted tuna sandwich with cheese" and "grilled tuna sandwich" are a kind of tuna sandwich. Furthermore, food entries often contain brand/restaurant name and serving size, e.g., "Chili's - Santa Fe Chicken Salad, 3 cups". Therefore, it is problematic to use the original text of the food entries in the analysis as two virtually identical entries might appear to be textually different.

To mitigate the problem, we represented each food entry as a set of salient features. A salient feature was extracted from the food entry text by matching word tokens of the food entry text with food-specific concepts, such as ingredients, preparation methods, etc. For this task, we manually built a food taxonomy[11] consisting of main categories, subcategories, and entities (leaf nodes). For each main category in the taxonomy, we find the maximal match between taxonomic entities in that category and the food entry text. After a match was found, we created a salient feature by concatenating the corresponding main category, subcategory, and entity (in "main category:subcategory:entity" format) and added the term to the set of salient features. We found that the procedure was effective enough in removing most noises from the original food entry text. For example, after preprocessing, the food entry "McDonald's - premium sweet chili chicken Wrap (grilled), 1 burger (200g)" will be represented as a set of 3 salient features {staple foods:wheat:wrap, meats:poultry:chicken, preparation methods:dry heat:grill}. As a result, the number of unique food entries in the dataset was significantly reduced from 1.2 million to 71.7 thousand entries. About 10% of food entries could not be matched to any entities and were subsequently discarded. Table 1 summarizes the dataset[2] after the preprocessing steps.

3.3 Building the Food-Context Matrix

Finally, given the processed food diary data, we computed PMI_{sig} [2] for each pair of food entry $f \in V_f$ and context $c \in V_c$, where c is another food entry occurred in the same meal as f, and built the food-context matrix M with 22,804 rows ($|V_f|$) and 63,653 columns ($|V_c|$). Then, we applied LSA to the matrix M to get the low-dimensional representation M_k. Typically, the value of k is in the [500, 5000] range. In this work, we set $k = 500$.

4. EXPERIMENTS

4.1 Constructing the Evaluation Dataset

We designed the evaluation of the food substitution extraction methods: PPMI matrix and SVD, as a top-k substitute ranking task. To obtain human judgements of food substitutes, we used CrowdFlower[3], an online crowdsourcing service. First, we randomly chose 100 food entries containing ingredients from major protein groups (i.e., meats, beans and legumes, and nuts and seeds) to be used as target queries. Next, we generated a ranked list of top-10 food substitute candidates for each target query using each method. This resulted in 2,000 food substitute pairs (1,000 for each method) to be labeled by CrowdFlower workers. Then, we instructed each worker to rate how likely they agree that each food pair is a suitable substitute of each other on 7-point Likert scale responses from 1 (strongly disagree) to 7 (strongly agree). Each food pair was judged by 3 workers. For quality control, 57 test questions created by the first author were used as ground truths to filter out low-quality workers. Cohen's Kappa between the workers' labels and the ground truth labels was 0.87, indicating strong agreement.

[2]https://goo.gl/Hkyi5w

[3]http://www.crowdflower.com/

Method	prec@1	prec@10	MAP	NDCG
PPMI Matrix	0.75	0.777	0.826	0.811
SVD	0.77	0.777	0.823	0.772

Table 2: Performance of the each method ($\tau = 3$).

Method	prec@1	prec@10	MAP
PPMI Matrix	0.58	0.567	0.673
SVD	0.69	0.696	0.755

Table 3: Performance of the each method ($\tau = 4$). Since NDCGs are not affected by the threshold, they are omitted from the table.

4.2 Evaluation Metrics

We employed 3 metrics used in standard evaluation of ranked lists in information retrieval: precision at k where $k = 1$ and $k = 10$, mean average precision (MAP), and normalized discounted cumulative gain (NDCG). To obtain the ground truth judgement for each food pair, we simply took the average of all ratings given by workers to the food pair. Since prec@k and MAP require binary judgements, we experimented with two binary judgement threshold τ values. Particularly, we were interested in comparing the performances when average ratings were greater than 3 (i.e., at least 'not disagree') or greater than 4 (i.e., at least 'slightly agree'). Any food pairs whose average ratings satisfying the binary judgement threshold are considered a true substitute pair. Lastly, all metrics have values in the [0, 1] range.

5. RESULTS & DISCUSSION

Table 2 shows the performance of PPMI matrix and SVD on the food substitute ranking task given $\tau = 3$. Overall, the results show that the vector space models can be effectively applied to extract food substitutes from food diaries. Both methods are equally good at identifying food substitutes according to prec@1, prec@10, and MAP. Next, with a more stringent threshold ($\tau = 4$), SVD greatly outperforms PPMI matrix according to prec@1 (+18.97%), prec@10 (+22.75%), and MAP (+12.22%). This is not surprising as SVD has shown to improve the similarity measurements in similar NLP tasks [8]. Interestingly, PPMI matrix is slightly better than SVD at generating ideal ranked lists of food substitutes according to NDCG (+5%). Examples of top-10 substitutes for the food entry "Tim Bacon - Bacon, 1 slices (54g)" extracted by PPMI matrix are shown in Table 4. Evidently, the algorithm ranked food entries containing processed meats (e.g., sausages and bacon) as suitable substitutes for bacon slices higher than other protein groups.

Food entry
Homemade - Sausage Balls, 8 -inch Ball
Sainsburys - Smoked Streaky Bacon Rashers, 2 rasher
Kroger - Traditional Cut Bacon, 2 slices
Pork Sausage, Spicy - Natures Promise, 1 Link
Pork - Cured, bacon, cooked, pan-fried, 2 slice cooked
Leidy's - Maple Glazed Premium Sliced Bacon, 1 Strips
Oscar Mayer - Turkey Bacon, 1 slice
Unknown - 2 Rasher of Grilled Bacon, 70 g
Bacon - Bacon Slices-oven Baked, 4 oven baked
Hormel - Black Label Bacon Original, 2 Pan Fried Slices

Table 4: Top-10 food entries identified as substitutes for "Tim Bacon - Bacon, 1 slices (54g)" extracted by PPMI matrix.

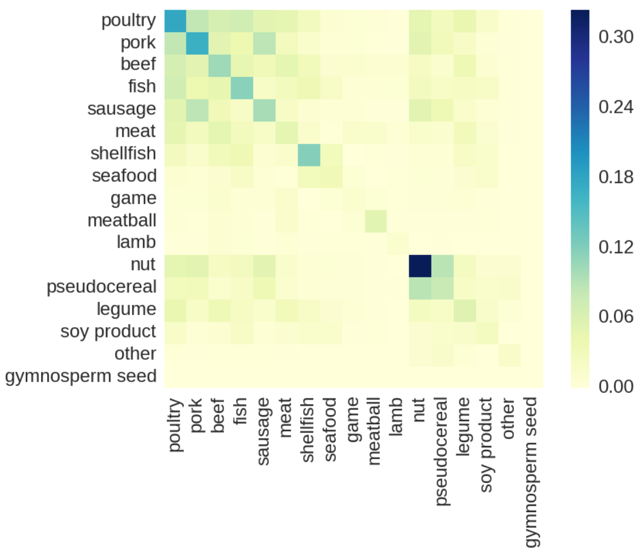

Figure 2: Normalized co-occurrences of substitutes.

Our study also shed light on the overall dietary behavior of the MFP users. Figure 2 summarizes the normalized co-occurrences of food subcategories from 19,040 substitute pairs containing ingredients from the major protein groups. Each cell represents the Jaccard normalization of co-occurrence of two food subcategories; the darker the cell color, the greater the co-occurrence. As we can see, most substitutions were between foods in the poultry subcategory, such as chicken and turkey, and other subcategories. Next, foods containing nuts were mostly substituted with nuts and other plant-based proteins. Lastly, plant-based proteins were rarely consumed in place of meat-based proteins.

6. CONCLUSION

This study investigated the task of extracting food substitutes from the self-reported food consumption data. Our approach is based on the assumption that foods consumed in similar contexts are more likely to be a suitable substitute of each other. We applied the vector space models of semantics, commonly used in NLP, to identify food substitutes and created ground truth judgements of 2,000 food substitute pairs to evaluate the effectiveness of the methods. The experiments showed promising results. Our work is not without some limitations. First, because the majority of our CrowdFlower workers were from countries outside the US (i.e., Indonesia, Venezuela, and India), their judgements could be affected by cultural biases when labeling food consumption data created by users in the US. Next, our data preprocessing steps can be further improved. As previously discussed, about 10% of all food entries had to be discarded due to the lack of salient features. For future work, we plan to experiment with other dense representation methods such as neural embeddings, incorporate higher-order co-occurrence and other contextual information, and identify "personalized substitutes". Lastly, to suggest "similar but healthier" options, we would also consider quantifying the healthfulness of foods through nutrient profiling.

7. ACKNOWLEDGEMENTS

This work is supported by the National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office.

8. REFERENCES

[1] C. Boscarino, N. J. J. P. Koenderink, V. Nedović, and J. L. Top. Automatic extraction of ingredient's substitutes. In *Proceedings of UbiComp '14 Adjunct*, pages 559–564, 2014.

[2] O. P. Damani and S. Ghonge. Appropriately Incorporating Statistical Significance in PMI. In *Proceedings of EMNLP 2013*, pages 163–169, 2013.

[3] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[4] J. Freyne and S. Berkovsky. Intelligent food planning: Personalized recipe recommendation. In *Proceedings of IUI '10*, IUI '10, pages 321–324, 2010.

[5] M. Ge, M. Elahi, I. Fernaández-Tobías, F. Ricci, and D. Massimo. Using Tags and Latent Factors in a Food Recommender System. In *Proceedings of DH '15*, pages 105–112, 2015.

[6] M. Harvey, B. Ludwig, and D. Elsweiler. You Are What You Eat: Learning User Tastes for Rating Prediction. In *Proceedings of SPIRE 2013*, pages 153–164, 2013.

[7] P. D. Howell, L. D. Martin, H. Salehian, C. Lee, K. M. Eastman, and J. Kim. Analyzing Taste Preferences From Crowdsourced Food Entries. In *Proceedings of DH '16*, pages 131–140, 2016.

[8] O. Levy, Y. Goldberg, and I. Dagan. Improving Distributional Similarity with Lessons Learned from Word Embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225, 2015.

[9] C.-Y. Teng, Y.-R. Lin, and L. A. Adamic. Recipe recommendation using ingredient networks. In *Proceedings of WebSci '12*, pages 298–307, 2012.

[10] P. D. Turney and P. Pantel. From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 37:141–188, 2010.

[11] I. Weber and P. Achananuparp. Insights from machine-learned diet success prediction. In *Proceedings of Pacific Symposium on Biocomputing (PSB)*, 2016.