Investigating substitutability of food items in consumption data

Sema Akkoyunlu
UMR MIA-Paris, AgroParisTech, INRA
Université Paris-Saclay
Paris, France
sema.akkoyunlu@agroparistech.fr

Cristina Manfredotti
UMR MIA-Paris, AgroParisTech, INRA
Université Paris-Saclay
Paris, France
cristina.manfredotti@agroparistech.fr

Antoine Cornuéjols
UMR MIA-Paris, AgroParisTech, INRA
Université Paris-Saclay
Paris, France
antoine.cornuejols@agroparistech.fr

Nicolas Darcel
UMR PNCA, AgroParisTech, INRA
Université Paris-Saclay
Paris, France
nicolas.darcel@agroparistech.fr

Fabien Delaere
Danone Nutricia Research
Palaiseau, France
fabien.delaere@danone.com

ABSTRACT

Food based dietary guidelines are insufficiently followed by consumers. One of the principal explanations of this failure is that they are too general and do not take into account eating habits. Providing personalized dietary recommendations via nutrition recommender system can hence help people improve their eating habits. Understanding eating habits is a keystone in order to build a context aware recommender system that delivers personalized dietary recommendations. As a first step towards this goal, we explore food relationships on real-world data using the INCA 2 dataset, a French consumption survey. We particularly focus on extracting food substitutions, i.e food items that can replace each other. We consider that two food items can be substituted if they are consumed during similar contexts. We define the context in the nutrition field and we introduce a measure of substitutability between food items based on consumption data that encodes the context.

CCS CONCEPTS

• **Information systems** → **Information extraction**; **Recommender systems**;

KEYWORDS

recommender system, substitution, food consumption, nutrition

ACM Reference format:
Sema Akkoyunlu, Cristina Manfredotti, Antoine Cornuéjols, Nicolas Darcel, and Fabien Delaere. 2017. Investigating substitutability of food items in consumption data. In *Proceedings of the Second International Workshop on Health Recommender Systems co-located with ACM RecSys 2017, Como, Italy, August 2017 (HealthRecSys'17),* 5 pages.

1 INTRODUCTION

Nutritional quality of diets is proven to be an important factor in health dysfunctions. The risk of developing modern chronic diseases such as cardiovascular diseases, obesity or diabetes is linked to unhealthy eating habits [13].

In order to promote healthy and sustainable diet and prevent chronic diseases, dietary guidelines targeted to the general population are produced by public health agencies. However, it has been noted that the compliance to the guidelines is usually low although the awareness concerning the food based dietary recommendations is rather good [9]. Nutrition related knowledge does not imply adherence to dietary guidelines. Several causes explain this phenomenon: cultural and personal preferences, difficulty of implementing dietary changes, availability of food items [12]. Nutritionists stress the fact that it is crucial to understand consumers' behaviours in order to make practical food-based recommendations because making changes is challenging [4].

One fair assumption is that people are more likely to follow recommendations if these are acceptable from their point of view. We hypothesize that the user acceptance is a prerequisite for the compliance and could be improved by producing user-tailored recommendations that take into account dietary habits. On the long term, our objective is to build a nutrition recommender system taking into account dietary habits in order to encourage people toward healthier alternatives with high compliance.

In food related recommender systems, the recommended items are recipes [6], [8] or food items themselves [7]. We rather want to build a food item based recommender system that delivers message such as "instead of eating X, eat Y". In order to deliver relevant recommendations, it is important for a recommender systems to know substitutability relationships between items [11], [14]. This is important in food recommender systems as well.

Moreover, it has been shown that context-aware recommender systems produce better recommendations than recommender systems that do not take into account the context [2]. In order to extract meaningful relationships between food items, in our model we consider contextual information. To the best of our knowledge, one study tackled the subject of food substitutability based on real-world consumption data [1]. However, they do not take into account contextual information such as the type of meal where substitutability relationships can be highly different.

In this paper, we specifically investigate food substitutability. To do that, we define the concept of *dietary context* as the set of food items a food is consumed with and the concept of *food intake* *context* as the setting of food consumption. Our intuition is that two food items are substitutable if they are consumed in similar dietary contexts and that substitutability differs according to the food intake context.

The rest of the paper is organized as follows. Section 2 describes our methodology. Section 3 reports the results. Finally in section 4, we discuss our results and present our future perspectives.

2 OUR APPROACH

2.1 Notations and problem statement

Let $\mathcal{X}$ be the set of food items. A meal is a collection of food items consumed at the same timeframe. For instance, *{coffee, bread, jam, juice}* is a meal. The meal database DB is the set of all meals. Let us denote $DB_{breakfast}$ the database of breakfasts and DB_{lunch} the database of lunches.

Our objective is to mine food pair substitutability applied by consumers when they compose their meals. Given a database of meals, we want to extract substitutability relationships based on the way people consume food. No nutritional information is used during this process. Instead, contextual information is used in order to extract meaningful substitutability relationships.

2.2 Defining Context

The notion of context is quite complex and difficult to define universally. In the field of recommender systems, the context is usually defined according to the field of application of the system.

In the nutrition field, we define two types of contexts: the dietary context and the food intake context. The **dietary context** of a food item x is the set of food items c with which x is consumed. For instance, in the meal *{coffee, bread, jam, juice}*, the dietary context of *{coffee}* is *{bread, jam, juice}*. We think that the dietary context is fundamental when seeking substitutability of food items because the way people compose their meals is intrinsically dependent on the relationships between the items.

The **food intake context** is defined as the set of all variables such as the type of the meal (breakfast, lunch, dinner, snack), the location (home, workplace, restaurant), the participants (family, friend, coworkers, alone). This corresponds to the notion of context usually used in context-aware recommender systems [2].

There are three paradigms for incorporating context in recommender systems : contextual pre-filtering, contextual post-filtering and contextual modelling [2]. Contextual pre (post)-filtering consists in splitting the dataset according to contextual variables before (after) applying algorithms. Contextual modelling consists in incorporating contextual information in the algorithm. In our framework, dietary context is used in order to model substitutability whereas the food intake context is used for contextual pre-filtering.

Our objective is to investigate substitutability among food items based on the assumption that two food items are highly substitutable if they are consumed in similar dietary contexts and in the same intake context.

Investigating all possible dietary contexts of a food item is computationally expensive because the number of possible dietary context is exponential in the number of food items and the length of the dietary context. The number of interesting contexts is actually limited by the characteristics of the available data. Instead of investigating all the dietary contexts of a food item, we decided to explore collections of meals that differ only by one item. We define the **dietary context** of a meal database c as the intersection of a set of meals S_m such that :

$$len(c) = \max_{x \in S_m} (len(x)) - 1 \tag{1}$$

Let us define the **substitutable set** S_c associated to a dietary context c as the set of food items such that the context c plus one item of S_c can be effectively consumed together. For instance, the substitutable set of the dietary context $c = \{bread, jam, juice\}$ might be $S_c = \{coffee, tea, yogurt\}$.

2.3 Mining substitutable items

To efficiently retrieve interesting sets of dietary contexts and their substitutable set, in this paper, we propose an approach based on graph mining techniques. Let us denote the meal graph $G = (V, E)$ where V is the set of nodes representing meals from the database and E is the set of edges such that two nodes are connected if there is at most one item that changes between them. A meal should appear at least once in the database in order to appear as a node in the graph. Figure 1 is a simple illustration of a meal network.

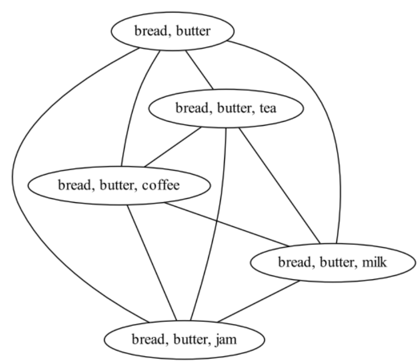

Figure 1: Example of a simple meal network

Designed in this way, the nodes of the substitutable set of a dietary context are adjacent. They form a sub-graph that is completely connected. Such an object is called a clique in graph mining. More specifically, the nodes form a maximal clique. A maximal clique is a clique to which another node cannot be added. In our setting, discovering substitutable sets is similar to mining maximal cliques in a graph. In this paper we use the algorithm of Bron-Kerbosh [5] to search for maximal cliques.

All discovered maximal cliques are not cliques that are interesting for our study. We want cliques such that the size of the intersection of the nodes is a dietary context as defined above. We denote these cliques as **substitutable cliques**. However, we may encounter cliques as in Figure 2. In this case, the intersection of the nodes is $\{A\}$ and we cannot derive a substitutable set from this clique.

To avoid retrieving uninteresting cliques, we apply Algorithm 1 that filters out substitutable cliques.

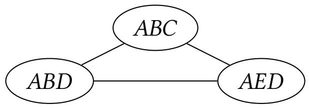

Figure 2: Example of an uninteresting clique

Algorithm 1 Find substitutable clique

```
function isSubclique(clique)
    context = getContext(clique)
    lenmax = max(len(x) for x in clique)
    if lenmax - len(context) = 1 then
        return True
    else
        return False
```

For instance, when we apply our algorithm to the example of Figure 1, we get that this graph is a maximal clique and a substitutable clique more particularly. The context is *{bread, butter}* and the substitutable set associated to this context is *{coffee, tea, milk, jam, nothing}*. In this particular case, it is possible to substitute an item by nothing because $\{bread, butter\}$ can be consumed as such.

2.4 Computing a substitutability score

Substitutability is not a binary relationship because there are different degrees of substitutability. Moreover, if two items are consumed together, they are less substitutable because they might be associated. Therefore, we need a function to quantify the relationship of substitutability that incorporates the possibility of associativity. Our hypothesis is that two items are highly substitutable if they are consumed in similar dietary contexts.

We want to compute a substitutability score such as :

(1) Two items are highly substitutable if they are consumed in similar contexts.
(2) Two items are less substitutable if they are consumed together.
(3) Substitutability is a symmetrical relationship.

Let us denote, for an item x, the context set C_x as the set of dietary contexts in which x is a substitutable item. If the cardinality of C_x denoted as $|C_x|$ is high, then x is substitutable in many dietary contexts.

For two items x and y, the condition (1) is described by the intersection of C_x and C_y. If $|C_x \cap C_y|$ is high, then x and y are consumed in similar contexts.

We denote $A_{x:y}$ the set of contexts of x where y appears :

$$A_{x:y} = \{c \in C_x | y \in c\} \tag{2}$$

The cardinality of $A_{x:y}$ denotes how y is associated to x.

Taking into account these considerations, we propose the **substitutability score** inspired by the Jaccard index [10]:

$$f(x,y) = \frac{|C_x \cap C_y|}{|C_x \cup C_y| + |A_{x:y}| + |A_{y:x}|} \tag{3}$$

The score equals 1 when x and y appear in exactly the same contexts and $A_{x:y} = A_{y:x} = \emptyset$. If x and y are never consumed in the same context then the score equals 0. The higher $|A_{x:y}| + |A_{y:x}|$ is, the higher the association of x and y is and the lower the score is.

3 EXPERIMENTS

3.1 The INCA 2 database

The French dataset INCA 2[1] is the result of a survey conducted during 2006-2007 about individual food consumption. Individual 7-day food diaries are reported for 2624 adults and 1455 children over several months taking into account possible seasonality in eating habits. A day is composed of three main meals : breakfast, lunch and dinner. The moments in between are denoted as snacking. For the main meals, the location (home, work, school, outdoor) and the companion (family, friends, coworkers, alone) are registered.

The 1280 food entries are organized in 44 groups and 110 subgroups of food items. We chose to consider the medium level of hierarchy in order to capture substitution relationships inter-groups and intra-groups.

Only adults are considered in this paper. All meals are gathered in a meal database DB_{meals} regardless of the type of meal. The database can be split according to contextual information in order to get better results [3]. We compare the results of our methodology on three datasets : $DB_{breakfastlunch}$, $DB_{breakfast}$ and DB_{lunch}.

3.2 Results

Applying our algorithm on $DB_{breakfast}$ yields 2368 contexts. Some of these and their substitutable sets are given in Table 1. Our results are coherent. For example, either bread, rusk or viennoiserie can be consumed for breakfast with coffee, sugar and water.

Context	Substitutable set
coffee, sugar, water, butter	bread rusk viennoiserie
tea/infusions, donuts	yogurt sugar jam/honey nothing

Table 1: Results of context and substitutable set retrieval for breakfasts

We applied our algorithm to the three datasets. The results are reported in Table 2. We can see that we can obtain inter-group substitutions such as *{potatoes ⇒ green beans}* but also intra-group substitutions as *{bread ⇒ rusk}*.

The substitutions proposed are consistent with regards to eating habits. Substitutes of drinks are also drinks : the substitutes of *coffee* are *tea*, *cocoa* and *chicory*. It is also the case for spreadable food items : the substitutes for butter for breakfast are spreadable items. No semantic information describing how a food item can be eaten is available in the dataset and yet considering the dietary context helps us retrieving this kind of information.

Substitutions between food items of the same nutritional food groups are found. For instance, the substitutes for *potatoes* are *pasta* and *rice*: they all contain starches.

[1] https://www.data.gouv.fr/fr/datasets/donnees-de-consommations-et-habitudes-alimentaires-de-letude-inca-2-3/

	Breakfast and lunch		Breakfast		Lunch	
Food Item	Substitute item (ordered by score)	Score	Substitute item (ordered by score)	Score	Substitute item (ordered by score)	Score
Bread	Rusk	0.2234	Rusk	0.3716	Fruits	0.0497
	Viennoiserie	0.1359	Viennoiserie	0.2010	Yogurt	0.0490
	Cakes	0.0745	Cakes	0.1243	Potatoes	0.0468
Coffee	Tea	0.2799	Tea	0.4219	Sodas	0.065
	Cocoa	0.1729	Chicory	0.2550	Yogurt	0.0642
	Chicory	0.1486	Cocoa	0.2255	Fruits	0.0633
Tea	Coffee	0.2799	Coffee	0.4219	Cakes	0.0536
	Cocoa	0.1721	Chicory	0.1965	Viennoiserie	0.0417
	Chicory	0.1289	Cocoa	0.1462	Coffee	0.0412
Cocoa	Chicory	0.2171	Chicory	0.2211	Cereal bars	0.25
	Coffee	0.1729	Coffee	0.2077	Preprocessed vegetables	0.0526
	Tea	0.1289	Tea	0.1965	Hamburgers	0.0256
Butter	Margarine	0.2413	Margarine	0.4030	Margarine	0.0602
	Honey/jam	0.0924	Chocolate spread	0.1240	Fruits	0.0431
	Chocolate spread	0.0786	Honey/jam	0.1175	Sauces	0.0431
Milk	Juice	0.1409	Yogurt	0.1815	Doughnut	0.0869
	Yogurt	0.1264	Juice	0.1504	Other milk	0.0666
	Sugar	0.1089	Tap water	0.1361	Milk in powder	0.0625
Wine	Sodas	0.0814			Sodas	0.0860
	Beer	0.0704	/	/	Tap water	0.0755
	Tap water	0.0412			Beer	0.0746
Pizza	Sandwich baguette	0.2429			Sandwiches baguette	0.2810
	Other sandwiches	0.1729	/	/	Other sandwiches	0.2177
	Meals with pasta or potatoes	0.1513			Meal with pasta or potatoes	0.1658
Potatoes	Pasta	0.1111			Pasta	0.1142
	Green beans	0.0922	/	/	Green beans	0.0941
	Rice	0.0602			Rice	0.0616

Table 2: Top 3 substitutable items for several items for breakfast and lunch

4 DISCUSSION AND CONCLUSIONS

We proposed a score of substitutability based on consumption data that can be used in a recommender system together with other scores such as a nutritional score that takes into account the nutritional contribution of the substitution and a user preference score. The substitutability score is based on the assumption that two items are substitutable if they are consumed in similar contexts. Preliminary results on the INCA2 dataset show that this assumption helps retrieving substitutability relationships based on consumption data.

When we split the dataset according to the contextual variable "type of meal", the substitutes and the scores are different. *Coffee* can be substituted by *tea*, *chicory* and *coffee* for breakfast whereas for lunch, it can be substituted by *sodas*, *yogurt* and *fruits*. Food items are consumed differently according to the type of meal. The relationship of substitutability is therefore different too.

Difference of scale in scores is noted according to the type of meal. It may be due to the fact that the diversity of food items consumed during lunch is higher than during breakfast. A rescaling factor based on the diversity of the type of meal can be introduced. As future work we plan to investigate this aspect and implement the nutritional score and the user preference related score.

5 ACKNOWLEDGEMENT

This study was funded by Danone Nutricia Research.

REFERENCES

[1] Achananuparp, P., and Weber, I. Extracting food substitutes from food diary via distributional similarity. *CoRR abs/1607.08807* (2016).

[2] Adomavicius, G., and Tuzhilin, A. Context-aware recommender systems. In *Recommender Systems Handbook*, F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, Eds. Springer US, 2011, pp. 217–253.

[3] Baltrunas, L., and Ricci, F. Context-based splitting of item ratings in collaborative filtering. In *Proceedings of the Third ACM Conference on Recommender Systems* (New York, NY, USA, 2009), RecSys '09, ACM, pp. 245–248.

[4] Bier, D. M., Derelian, D., German, J. B., Katz, D. L., Pate, R. R., and Thompson, K. M. Improving compliance with dietary recommendations. *Nutrition Today 43*, 5 (sep 2008), 180–187.

[5] Bron, C., and Kerbosch, J. Algorithm 457: Finding all cliques of an undirected graph. *Commun. ACM 16*, 9 (Sept. 1973), 575–577.

[6] Freyne, J., and Berkovsky, S. Intelligent food planning: personalized recipe recommendation. In *Proceedings of the 15th International Conference on Intelligent User Interfaces, IUI 2010, Hong Kong, China, February 7-10, 2010* (2010), pp. 321–324.

[7] Ge, M., Elahi, M., Fernaández-Tobías, I., Ricci, F., and Massimo, D. Using tags and latent factors in a food recommender system. In *Proceedings of the 5th International Conference on Digital Health 2015* (New York, NY, USA, 2015), DH '15, ACM, pp. 105–112.

[8] Harvey, M., Ludwig, B., and Elsweiler, D. You are what you eat: Learning user tastes for rating prediction. In *String Processing and Information Retrieval - 20th International Symposium, SPIRE 2013, Jerusalem, Israel, October 7-9, 2013, Proceedings* (2013), pp. 153–164.

[9] Ivens, B. J., and Smith Edge, M. Translating the Dietary Guidelines to Promote

Behavior Change: Perspectives from the Food and Nutrition Science Solutions Joint Task Force. *J Acad Nutr Diet 116*, 10 (Oct 2016), 1697–1702.

[10] Jaccard, P. The distribution of the flora in the alpine zone.1. *New Phytologist 11*, 2 (1912), 37–50.

[11] McAuley, J. J., Pandey, R., and Leskovec, J. Inferring networks of substitutable and complementary products. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015* (2015), pp. 785–794.

[12] Webb, D., and Byrd-Bredbenner, C. Overcoming consumer inertia to dietary guidance. *Adv Nutr 6*, 4 (Jul 2015), 391–396.

[13] World Health Organization. Diet, nutrition and the prevention of chronic diseases: report of a joint who/fao expert consultation. Tech. rep., 2003.

[14] Zheng, J., Wu, X., Niu, J., and Bolivar, A. Substitutes or complements: another step forward in recommendations. In *Proceedings 10th ACM Conference on Electronic Commerce (EC-2009), Stanford, California, USA, July 6–10, 2009* (2009), pp. 139–146.