Machine Learning in Network Science

So…what do you wanna eat?

Project report

Aiza AVILA
DSBA M2
ESSEC Business School and Centrale Supélec, France
aiza.avila-canibe@student-cs.fr

Aline HELBURG
DSBA M2
ESSEC Business School and Centrale Supélec, France
aline.helburg@student-cs.fr

Joshua Jehau FAN
DSBA M2
ESSEC Business School and Centrale Supélec, France
joshua.fan@student-cs.fr

Xiaoyan HONG
DSBA M2
ESSEC Business School and Centrale Supélec, France
xiaoyan.hong@student-cs.fr

ABSTRACT

In this project, we propose a machine learning-based approach to explore the similarity between the ingredients of different cuisines and determine their pairing of flavors. We construct ingredient networks for each cuisine by representing ingredients as nodes and co-occurrences as edges. We employ machine learning's algorithm to model the relationships between the ingredients, and use it to generate feature importance scores for each ingredient. By analyzing these scores, we identify the key flavor components of each cuisine, and compare them to find similarities between different cuisines. We then use this information to recommend dishes based on their similarity in terms of flavor pairing. Our approach proves to be effective in identifying flavor similarities between cuisines, and can be used to recommend dishes that are likely to be enjoyed by individuals based on their culinary preferences. Our best-performing cuisine classifier was Support Vector Machine (SVM) with an accuracy of 70.1%. This approach also has the potential to enhance the culinary experience for individuals by providing tailored dish recommendations based on their flavor preferences.

CCS CONCEPTS

• Network Science • Machine Learning • Graphical Network

KEYWORDS

Food Computing, Food Pairing, Flavor, Ingredient Network, Similarity

1 Introduction

Machine learning and network science have both emerged as important fields in recent years, and their combination has led to new and exciting applications. The ability of machine learning algorithms to learn patterns and relationships from large datasets, combined with the power of network analysis to model complex systems, has led to a surge of interest in this interdisciplinary area of research. Machine learning techniques have been used in network science to analyze large-scale networks and uncover hidden patterns and structures. Their application scope is broad - from bioscience to social media and customers' recommendations. The ability to analyze complex systems and uncover hidden patterns has already led to important applications in fields such as healthcare, social media, and culinary arts. With continued research and development, this interdisciplinary field is sure to lead to even more exciting breakthroughs in the years to come. With this project, we aim to dig into one application that is of high interest to us, with a high significance and technically challenging.

As a result of our participation in the master's program, each of the students has been able to exchange points of view on different topics with our peers. And perhaps the most repeated has been food. Some of us miss the dishes of our homeland and have had to adapt to the ingredients available in France, making the culinary experience sometimes not 100% authentic. Having a shared feeling of longing, we have recreated regional dishes on multiple occasions, sharing experiences and flavors with our colleagues, and on more than one occasion we have been surprised by the similarities and differences we have in the use of ingredients. Motivated by this fact, we see in the analysis of ingredients a possibility to break down the flavors that unite us and those that make us unique.

2 Related work

The main scientific work on the flavor-ingredient network has been conducted in 2011 by Yong-Yeol Ahn et al. [1]. In their article, they developed a flavor network to depict the flavor compounds that are commonly found in various culinary ingredients. They found out that Western cuisines tend to use ingredient combinations that have a high number of shared flavor compounds. These cuisines commonly use ingredients that have similar molecular components in their recipes. However, this trend is not universal as it tends to break down when it comes to cuisines outside of these regions. In particular, East Asian and Southern European recipes tend to incorporate ingredients that do not share common flavor compounds, indicating that these styles of cooking are significantly different from each other in terms of the quantitative aspects of their ingredient combinations. Their network was based on the molecular level of food and garnered significant interest from the non-scientific press.

Other researchers also studied food pairing as part of network science. In 2012, Chun-Yuen Teng et al. [2] published an article that stated that the network they have developed provides information about which ingredients are suitable to be used together and which ones can be replaced to enhance the overall results. This network also enables one to anticipate which recipe among a pair of related recipes is more likely to receive a higher rating from its users.

In 2015, Jain et al. [3] focused their work on the study of food pairing between regional cuisines of India. Their article reveals that the more two ingredients share similar flavors, the less likely they are to be used together in that particular regional cuisine.

In 2015, a Mexican team [4] used social media as a methodology to research the relationship between foods and beer pairing. Image-based social medias like Instagram give key information about how users pair their foods and beverages.

In 2013, Varshney et al. studied the food pairing in Medieval Europe cuisine [5]. Medieval cuisine tends to share more compounds than the cuisine today.

3 Methodology and results

Our project is based on data from Yong-Yeol Ahn's work [1], including 1530 food ingredients, flavor compounds and 36,781 edges. We also have a database of 7,000 recipes.

3.1. Preprocessing

The first step was to disaggregate the recipe database of 7,000 elements to have a list of ingredients. This step involves breaking down each recipe in the database into a list of its constituent ingredients. For example, if a recipe calls for "pasta with tomato sauce and basil", this step would break it down into three ingredients: pasta, tomato sauce, and basil. We homogenize the ingredient list by performing a lemmatization, a regex and splitting the words. We also removed words that bring low information such as ‘low fat’, ‘gluten free’, etc.

Then we matched ingredients from the recipes to ingredients in a graph, and filtered out those with more than three missing matches. We made sure that the functions chosen also catch some spelling differences and make appropriate substitutions. Because recipes had a high variability, we standardized the ingredient names and made them consistent across the entire dataset. For example, the words : 'rib', 'chuck', 'sirloin' and 'steak' were replaced with ‘beef’. Finally, we choose to perform a function that makes additional substitutions for certain ingredients that were not initially matched correctly. Using the compound database, we matched each ingredient to its corresponding flavor and removed the ingredients that do not have a match in the flavor dataset.

We ended up with a dataset with columns for each ingredient, as well as a dataset projected into the ingredient space in the flavor network. Some recipes had to be discarded as their ingredients are not listed in the flavor network.

We perform weight scaling of ingredients depending on the recipe in which they are. This step involves assigning weights to each ingredient based on its quantity in the recipe. For example, if a recipe calls for 2 cups of pasta and 1 cup of tomato sauce, pasta would have a weight of 2 and tomato sauce would have a weight of 1. For this step, we used the TF-IDF "Term Frequency-Inverse Document Frequency" method. The TF-IDF method is used to normalize the flavor matrix to adjust for the relative importance of each compound in each recipe.

3.2. Backbone

In order to create the nodes and edges, we calculated all the combinations possible between the different ingredients. The nodes are the top 200 ingredients. The edges were created by counting the number of recipes where the ingredient lives. The ingredients with the highest degrees were the ones present in the highest number of recipes - here onions are in 18 205 recipes, followed by garlic with 17 465 recipes. The importance of each edge was given by the count of the pair (ingredient 1, ingredient 2): for example, the pair (almond, apricot) was counted 7 times which is less than the pairs (almond, anis) counted 12 times.

The one-sided ratio for each ingredient pair is determined by dividing the number of co-occurrences by the frequency of the source node.

Then the other types of weights are calculated using the ingredient-pair frequency-inverse recipe frequency (IF-IRF) method. Firstly, the recipe counts for each ingredient in the ingredient pairs are retrieved from the graph information. Then, the combined recipe count for each ingredient pair is calculated by summing the recipe counts of the two ingredients. The ingredient-pair frequency (IF) is computed as [1 + log10(1 + IF)) * log10(total recipes / recipe frequency]. The inverse recipe frequency (IRF) is obtained by taking the logarithm of the total number of recipes divided by the combined recipe count. The IF and IRF values are then multiplied to obtain the IF-IRF weight. Additionally, a log (base e) of IF-IRF is also calculated by replacing the log (base 10) with natural logarithm in the IF and IRF formulas.

The IF-IRF weight calculated with log (base e) is then used to extract the backbone graph. The threshold is set as 60% of the maximum weight, the edge of the pair which has a weight below the threshold will be removed, resulting in the extraction of a backbone graph that retains only edges with weights above the threshold.

The backbone graph is visualized as follows:

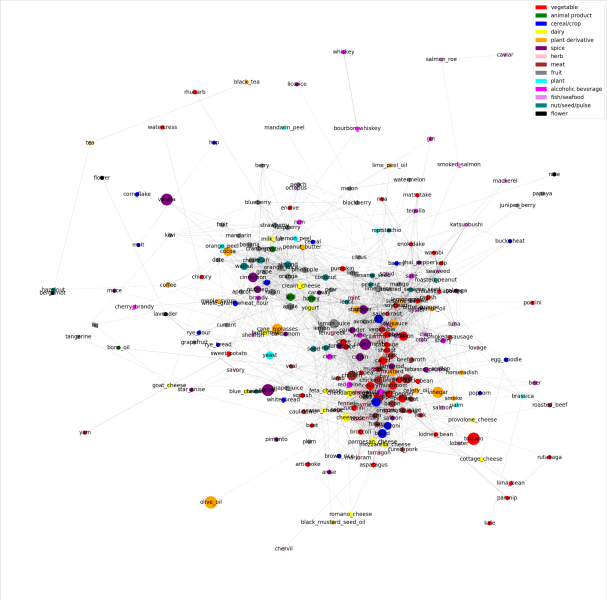

Figure 1. Backbone of the ingredient network

Each node denotes an ingredient. Two ingredients are connected if they share a significant number of flavor compounds. From the figure, we can see some commonly used ingredient types (for example: vegetables, spices, cereals) and the clustering relationship of some ingredients, we will further explore in the following parts. Here, we kept the backbone part to visualize the graph. We still used the full network in the rest of our work.

3.3. Cuisine classification

We classified the different cuisines based on the ingredient matrix and on the flavor matrix.

3.3.1. Cuisine classification with ingredients

We used ingredient information as features. We apply the following classifiers in order to classify the cuisines:

- Logistic Regression
- Support Vector Machine
- RandomForest
- Multinomial Naive Bayes
- XGBoost

Table 1. Evaluation of the performance of the classifiers

Classifier	Accuracy

Logistic Regression	69.3%
Support Vector Machine	**70.1%**
RandomForest	68.4%
Multinomial Naive Bayes	63.8%
XGBoost	68.8%

We also performed hyperparameter tuning in order to prevent overfitting and achieve the highest performance possible.

SVM is the classifier that allows the best classification of the recipes. Using ingredient information as features, we were able to classify recipes into regional cuisines with 71% accuracy.

3.3.2. Cuisine classification with flavor

Then, we used the flavors as features. We apply the following classifiers in order to classify the cuisines:

- Logistic Regression
- Support Vector Machine

Table 2. Evaluation of the performance of the classifiers

Classifier	Accuracy
Logistic Regression	**66.2%**
Support Vector Machine	65.2%

Using flavor profile as features, our classification accuracy using logistic regression is only 66%, suggesting more overlap in the flavor space.

3.4. Cuisine clustering

We performed t-SNE clustering and plotted on a dataframe of recipes based on their regional cuisine. t-SNE is commonly used for visualizing high-dimensional data in a lower-dimensional space (usually 2D or 3D). We choose this algorithm because it is a non-linear dimensionality reduction technique that aims to preserve the local structure of the high-dimensional data points in the lower-dimensional space.

We first separated the cuisines into two groups: cuisines that are alike and cuisines that are different. This first 'manual' clustering was performed in order to better visualize the clusters - as we have a high number of cuisines. This grouping was done based on their cultural background. For example, the cuisines that we assumed were alike are cuisines from North America (Mexican, Cuban, Hawaiian, etc).

3.4.1. Cuisines that are alike

In this group, we find the following cuisines: 'Southern & Soul Food, American', 'American', 'Southwestern, American, Mexican', 'Cajun & Creole, American', 'Southwestern, American', 'Cajun & Creole, Southern & Soul Food, American', 'Hawaiian, American', 'Chinese, Asian'.

We performed three measure distances: Jaccard, cosine and hamming. Jaccard and cosine have both poor performances: they struggled to differentiate the regional cuisines of North America - which may mean they share similarities.

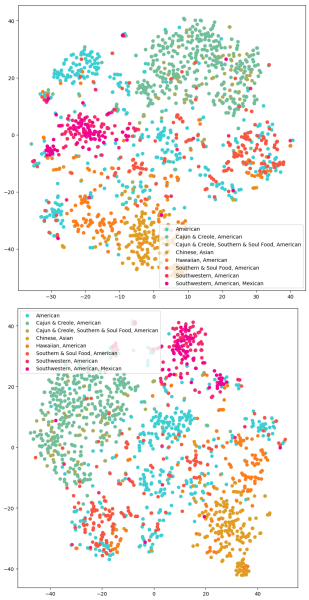

Figure 2. Representation of the cuisines based on the cosine distance (left) and the Jaccard distance (right)

The hamming distance performed better as we could visualize more defined clusters.

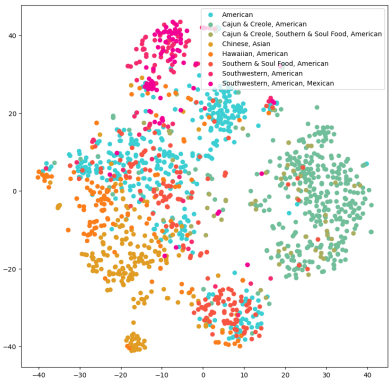

Figure 3. Representation of the cuisines based on the hamming distance

3.4.2. Cuisines that are different

In this group we find the following cuisines : 'Indian, Asian', 'Mexican', 'Greek', 'Cajun & Creole, American', 'Chinese, Asian', 'Italian' and 'Irish'.

3.4.2.1. Clustering on ingredients

In this group we find the following cuisines: 'Indian, Asian', 'Mexican', 'Greek', 'Cajun & Creole, American', 'Chinese, Asian', 'Italian' and 'Irish'. Such as the section before, we performed the three same distance measures.

With these new cuisines, we can better visualize the differentiation between each cuisine. However, it appears to be difficult to properly cluster the Irish cuisine (in dark orange) from the others. This could be explained by the fact that the Irish cuisine has an influence on other American cuisines - as the Irish immigrated to america in the 19th century.

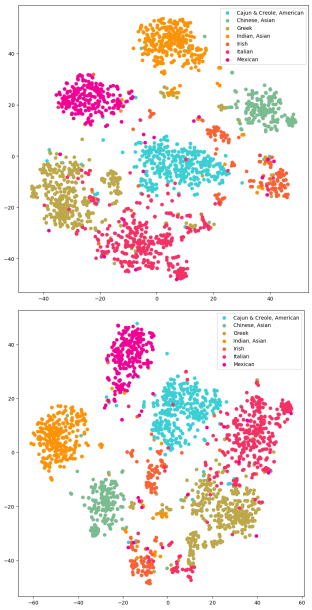

Figure 4. Representation of the cuisines based on the cosine distance (left) and the Jaccard distance (right)

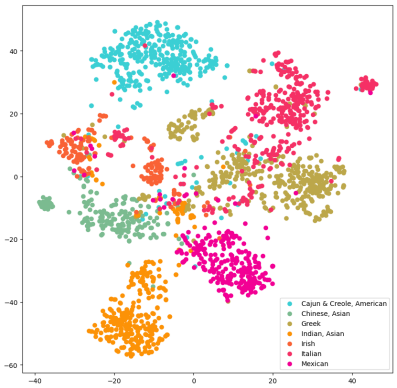

Figure 5. Representation of the cuisines based on the hamming distance

The hamming distance is struggling more with Greek and Italian cuisines (beige and red). It is also difficult to distinguish between Chinese, Indian and Mexican cuisines.

3.4.2.2. Clustering based on flavors

To perform clustering based on the flavor compounds, we used the following distance measure (as the three methods used before were not enough to distinguish between cuisines): hamming, Jaccard, cosine, euclidean, sqeuclidean and Minkowski.

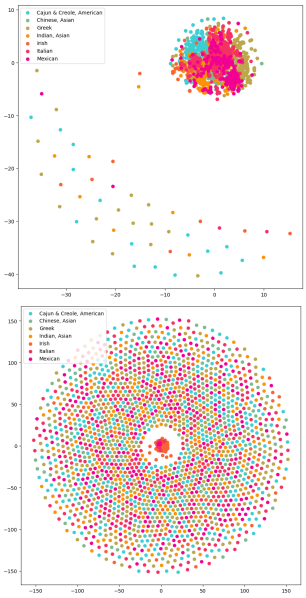

Figure 6. Clustering based on flavors using hamming distance (left) and Jaccard (right)

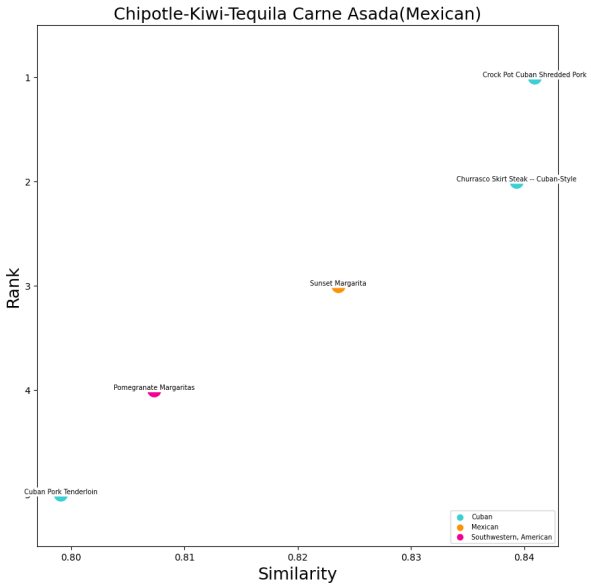

Figure 7. Top 5 dishes recommendations for the Mexican dish Chipotle Kiwi Tequila Carne Asada.

While it is beautiful to look at (especially for the Jaccard distance), the clustering based on flavor does not allow to distinguish between cuisines.
We can suggest that the flavors can not be clustered to a specific cuisine since there are ingredients shared through them all, like tomatoes and garlic.

3.5. Dish recommendation

Based on the ingredients present in a recipe, we built a function that returns the closest recipes based on the cuisine clustering. The closest recipe is calculated using the cosine similarity. As an example, if the user chooses as an input a Mexican dish, the function can return either a similar Mexican dish or a Cuban one. The user chooses if he prefers a similar dish from the same original cuisine or not.

Using the example given in the graph above, the closest recipe from Chipotle Kiwi Tequila Carne Asada is the crock pot cuban shredded pork (Cuban cuisine) with a similarity of 0.84 followed by another cuban dish, the Churrasco Skirt Steak with a similarity of 0.835.

4 Evaluation

Regarding the classification of the cuisines, we plotted the confusion matrix in order to better visualize the performance of our classification model. The rows represent the true labels and the columns represent the predicted labels. The diagonal of the confusion matrix shows the number of correct classifications, while the off-diagonal elements indicate the misclassifications.
We managed to see some similarities, for example, the Mediterranean cuisine and the Greek cuisine appear to share the same ingredients and thus ‘confuse’ our classification model.

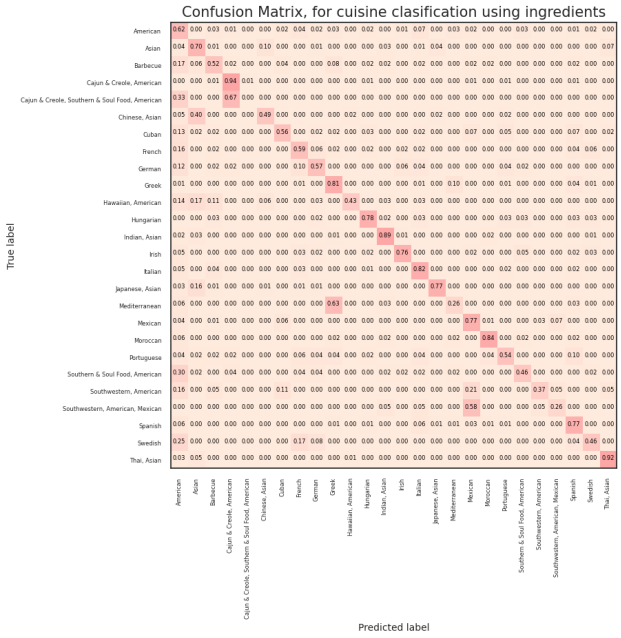

Figure 8. Confusion matrix after performing the SVM classifier

When analyzing the dish recommendation, we may face some strangeness. In the previous example, a similar dish of the Chipotle Kiwi Tequila Carne Asada is Sunset Margarita. The user needs to keep in mind that his dish recommendation is ingredient- based, and both recipes contain Tequila, orange juice, and lime juice. It is worth noting that the output of our dish recommendation should be treated cautiously - for obvious reasons. A solution that can be implemented is to add weights to ingredients that are more relevant than others. We performed a weight scaling based on the quantity but on the relevance. For example onions, garlic, olive oil should have lower weights since they are present in most recipes.

The flavor network does not provide information on the concentration of each compound - a compound is considered as present above a certain threshold. The concentration of each compound plays an important role in taste and flavor and thus in dish recommendation.

5 Conclusions

The proposed machine learning-based approach is effective in identifying and recommending dishes based on the similarity of flavors between cuisines, and has the potential to enhance the culinary experience for individuals by providing tailored dish recommendations based on their flavor preferences.

However, it is important to know the complexity of human taste perception, which is influenced by other sensory experiences, such as smells and temperature. Further research can also be conducted to explore the relationships between flavor compounds and cuisines, as well as incorporating cultural and geographical influences on cuisine to have a more comprehensive understanding of the diversity of cuisines worldwide. Additionally, we focused on traditional machine learning techniques in this research, exploring state-of-the-art deep learning models may further improve the performance of the approach.

REFERENCES

[1] Ahn, YY., Ahnert, S., Bagrow, J. et al. (2011) Flavor network and the principles of food pairing. Sci Rep 1, 196. https://doi.org/10.1038/srep00196

[2] Teng, C.-Y., Lin, Y.-R., & Adamic, L. A. (2012). Recipe recommendation using ingredient networks. In Proceedings of the 4th Annual ACM Web Science Conference (pp. 298-307). https://doi.org/10.1145/2380718.2380757

[3] Jain, A., N K, R., & Bagler, G. (2015). Analysis of Food Pairing in Regional Cuisines of India. PloS one, 10(10), e0139539. https://doi.org/10.1371/journal.pone.0139539

[4] Arellano-Covarrubias, A., Gómez-Corona, C., Varela, P., & Escalona-Buendía, H. B. (2019). Connecting flavors in social media: A cross cultural study with beer pairing. Food research international (Ottawa, Ont.), 115, 303–310. https://doi.org/10.1016/j.foodres.2018.12.004

[5] Varshney, K. R., Varshney, L. R., Wang, J., & Myers, D. (2013). Flavor pairing in medieval European cuisine: A study in cooking with dirty data. arXiv preprint arXiv:1307.7982. https://doi.org/10.48550/arXiv.1307.7982