Hierarchical Clustering of World Cuisines

1st Tript Sharma
Department of Mechanical Engineering
Delhi Technological University
New Delhi, India
triptsharma22@gmail.com

1st Utkarsh Upadhyay
Department of Electronics Engineering
Jamia Millia Islamia University
New Delhi, India
utkarshdhy@gmail.com

2nd Jushaan Kalra
Department of Computer Engineering
Delhi Technological University
New Delhi, India
jushaan18@gmail.com

3rd Sakshi Arora
Department of Computer Science
Indraprastha Institute of Information and Technology
New Delhi (IIIT-Delhi), India
sakshi18133@iiitd.ac.in

4th Saad Ahmad
Department of Computational Biology
Indraprastha Institute of Information and Technology
New Delhi (IIIT-Delhi), India
saad18409@iiitd.ac.in

4th Bhavay Aggarwal
Department of Computational Biology
Indraprastha Institute of Information and Technology
New Delhi (IIIT-Delhi), India
bhavay18384@iiitd.ac.in

Ganesh Bagler
Center for Computational Biology
Indraprastha Institute of Information and Technology
New Delhi (IIIT-Delhi), India
bagler@iiitd.ac.in

***Abstract*—Cultures across the world have evolved to have unique patterns despite shared ingredients and cooking techniques. Using data obtained from RecipeDB, an online resource for recipes, we extract patterns in 26 world cuisines and further probe for their inter-relatedness. By application of frequent itemset mining and ingredient authenticity we characterize the quintessential patterns in the cuisines and build a hierarchical tree of the world cuisines. This tree provides interesting insights into the evolution of cuisines and their geographical as well as historical relatedness.**

***Index Terms*—Hierarchical Clustering, Pattern Mining, Authenticity Correlation, Kmeans Clustering, Food Ontology**

I. Introduction

Cultures across the world have evolved diverse cooking practices over time. Although the underlying fundamentals of cooking remain the same, various factors including geography and climate have affected cooking styles. Cuisines from across the globe have thus acquired their signature styles. Each cuisine has interesting patterns that are inherent to it while sharing some common attributes with others. In this article, we characterize the unique features that typify every cuisine in an attempt to discern the footprint of food on human cultures and inter-relatedness of world cuisines.

II. Literature Survey

With increasing availability of recipes data, there has been much interest in data mining recipes data. One of the focus has been on defining recipe similarity. Attempts have been made to define similarity based on various elements of cooking recipes [15] and ingredients [7], [14].

Among other efforts in data mining recipes data have focused on food pairing phenomena in cuisines. Among one of the early studies, Shidochi et al [11] experimented with the possible replacements of ingredients in a recipe. Jain et al [8] investigated the phenomenon of food pairing which examines compatibility of two ingredients in a recipe in terms of their shared flavor compounds. This study investigated the food pairing phenomena in Indian recipes and proclaimed that spices form the basis of their food pairing. The work was extended by Singh et al [12] to analyze a much larger dataset encompassing 22 cuisines across the world and found interesting food pairing patterns in cuisines from across the world. Tuwani et al [13], on the other hand, considered culinary systems as a function of socio-cultural factors and presented computational models for cuisine evolution. An interesting work by Yokoi et al [16] calculated an ingredient associative metric called 'typicality value', giving out typical recurring ingredient patterns.

In this article, we indulge in pattern analysis in world-wide cuisines by way of association rule discovery and frequent pattern mining [1]. Going beyond the application of pattern mining techniques on cuisines in [10], we propose their use for frequent pattern mining of recipe data and cooking processes, utensils and ingredients for hierarchical clustering of cuisines.

III. Data Collection

Our analysis involved four types of information pertaining to traditional recipes, namely, recipes, ingredients, processes and utensils. A total of 118,071 recipes were obtained from various sources: AllRecipes, Food Network, Epicurious and TarlaDalal. RecipeDB [3], a structured compilation of recipes, was used as the primary source of information. All data are available at 'RecipeDB: A resource for exploring recipes'.

For each recipe, details such as its name and the list of ingredients and processes involved while cooking are

2473-3490/20/$31.00 ©2020 IEEE
DOI 10.1109/ICDEW49219.2020.00007

TABLE I
SIGNIFICANT PATTERNS MINED FROM CUISINES ACROSS THE WORLD

Region	Number of Recipes	Pattern	Support	Number of patterns
Australian	5,823	Butter	0.24	29
Belgian	1,060	Butter + salt	0.24	51
Canadian	6,700	Onion	0.20	31
Caribbean	3,026	Garlic Clove	0.24	32
Central American	460	Onion	0.30	38
Chinese and Mongolian	5,896	Soy sauce + add + heat	0.27	88
Deutschland	4,323	Onion	0.29	54
Eastern European	2,503	Cream	0.30	60
French	6,381	skillet	0.21	60
Greek	4,185	Olive Oil	0.40	43
Indian Subcontinent	6,464	Onion + add + heat + salt	0.22	119
Irish	2,532	Butter	0.32	41
Italian	16,582	Parmesan cheese	0.31	63
Japanese	2,041	Soy Sauce	0.45	45
Mexican	14,463	cilantro	0.25	33
Rest Africa	2,740	Onion + add + heat	0.20	51
South American	7,176	Onion + salt	0.21	62
Southeast Asian	1,940	Fish sauce	0.24	69
Spanish and Portuguese	2,844	Olive Oil	0.31	67
Thai	2,605	Fish sauce + add + heat	0.23	73
Korean	668	Soy sauce + sesame oil green onion + sesame oil	0.34 0.24	85
Middle Eastern	3,905	Salt + bowl Lemon Juice	0.22 0.22	46
Northern Africa	1,611	cumin + cinnamon cumin + olive oil cumin + Salt	0.21 0.22 0.22	134
Scandinavian	2,811	Butter + Salt Salt + Sugar	0.22 0.21	52
UK	4,401	Butter Oven	0.37 0.46	45
US	5,031	Bake + preheat+ oven + bowl Onion	0.22 0.25	67

available. Each recipe was treated as an unordered list of ingredients, processes and utensils. We integrated recipes from all the sources and grouped them into 26 distinct geo-cultural 'cuisines' while ensuring that each region had enough recipes attributed to it to distinguish it as a cuisine. Please refer to Table I for the list of regions. Due to insufficient information about the region for many recipes, they were aggregated together on the basis of their geological similarities with the prefix 'Rest'. For example, recipes without 'region' information belonging to Africa were put in 'Rest Africa' category.

The database consists of 20,280 unique ingredients, 268 unique processes and 69 unique utensils. The data are sparse in the list of utensils and 14,601 recipes don't have information regarding the preferred utensils required for cooking. An average recipe in a cuisine has $\sim$10 ingredients, $\sim$12 processes and $\sim$3 utensils. This is intuitive as too many ingredients would impede the success/propagation of a recipe, whereas too few would lead to it being modified easily. Thus recipes needs to maintain a balance between complexity and simplicity to survive successive iterations of evolution [13].

IV. PATTERN MINING

To investigate the ontology of food, we mined rules from the data to understand the patterns that are prevalent in a given cuisine. The methodology employed for mining patterns is explained in Section V. The mined patterns consist of ingredients, processes and utensils permutations that have a frequency greater than the defined threshold support. According to [1], [6], support represents the frequency with which the collection of items co-occur as a percentage of all transactions. A high support threshold represents high confidence in the pattern being mined, whereas with a low support threshold noise can creep into the mined patterns, leading to false identification of cuisine features. Hence, a trade off support of 20% was chosen as the threshold. Since the mined patterns are the most frequent ones, it is safe to say that most of the recipes follow the observed patterns and essentially define the cooking practices of the cuisine.

Among the patterns obtained in all the recipes for the Korean region as shown in Table I, the pattern "Soy Sauce + sesame oil" occurs with a support of 0.34, i.e. the pattern is found in 34% of all the recipes in the Korean region. Table I contains the topmost significant patterns in the 26 cuisines. The pattern depicts set of words occurring in a particular

recipe. The patterns mined are highly skewed, with most regions containing patterns having generic ingredients such as 'salt', 'onion' and processes such as 'add' and 'cook', which is justified as they have a high frequency among all cuisines and are fundamental to cooking in many cuisines.

V. Methodology

We implemented two approaches namely Frequent Itemset-based Hierarchical Clustering (FIHC) [5] and Authenticity based Clustering [2] to extract relationships between various cuisines of the world. The hypothesis is that some patterns which are common across a subset of cuisines would be found, which defines their 'similarity/closeness' with each other.

A. Frequent Pattern Mining

Frequent Itemset Mining refers to discovering interesting patterns in databases such as association rules from a set. Since we treat a recipe as a combination of ingredients, processes and utensils, it can be treated as an unordered set of these entities. For the frequent itemset mining, the FP-Growth Algorithm [6] was used as it is an efficient and scalable method for mining the complete set of frequent patterns by pattern fragment growth. The data extracted from RecipeDB was pre-processed to make it compatible with the input form of FP-Growth Algorithm. Ingredients, utensils and processes were concatenated and the FP-Growth Algorithm was applied. This approach was applied to all 26 regions present in the data extracted from RecipeDB. The support was kept at 0.2 so that the pattern was mined across a reasonable number of recipes.

B. Authenticity Based Clustering

We propose that a cuisine can be represented as a set of ingredients, process and utensils which can thus be utilized to define the relationships among the cuisines. Using the authenticity metric described in [2] we calculate the prevalence P_i^c of an item i in a cuisine c according to equations 1 as a function of number of recipes, n_i^c in a cuisine over total number of recipes in the dataset, N_C. This is used to calculate the authenticity of the item for a cuisine using equation 2.

$$P_i^c = n_i^c / N_C \tag{1}$$

$$p_i^c = P_i^c - (P_i^k)_{c \neq k} \tag{2}$$

In order to obtain the contribution of the item in uniquely identifying a cuisine, a relative prevalence matrix is created by subtracting the average prevalence of the item, say i for all cuisines from the prevalence for cuisine c. Accordingly, the most prevalent and least prevalent items in a cuisine can be identified. It should be noted that both the most prevalent and least prevalent items would contribute towards the culinary fingerprint of a cuisine as the former indicates the items having a relatively higher utility in the cuisine while the latter indicates items that are least used in the cuisine versus the rest of the world cuisines.

VI. Clustering Techniques

A. Hierarchical Clustering

Application of FP-Growth Algorithm on the prepared dataset results in 26 files, each containing patterns in a 'frozenset' along with their respective support to remove redundant patterns. These patterns were extracted from the 'frozenset' and appended together in a list in a sorted fashion. All the elements of this list are appended and converted into a string resulting in a 'string pattern'. All the 'string patterns' are compiled into a set resulting in unique set of patterns across all the 26 regions. Since the data is in string form and each element is a unique entity, it can be classified as a category. Therefore, unique set of 'string patterns' are fit for using Label Encoding (because the strings are categorical data) to get a transformer and the 'string patterns' in the rules are transformed using the derived transformer across all the regions. All the 'string patterns' in the rules from all the regions are appended in an array. The data from this array is thus vectorized to form a feature vector which is thus fed to the cluster as the linkage matrix.

Three different approaches were applied in order to cluster the linkage matrix data and to generate subsequent dendrograms. The linkage matrix is converted into a condensed distance matrix (pdist) in order to calculate the distance between all the cuisines based on the rules mined and is then fed into the hierarchical clustering model. To analyze the clusters we have used three distance metrics:

$$\text{Jaccard Distance} = \frac{c_i \cup c_j}{c_i \cap c_j} \tag{3}$$

$$\text{Cosine Distance} = \frac{c_i . c_j}{|c_i||c_j|} \tag{4}$$

$$\text{Euclidean Distance} = \sqrt{c_i^2 + c_j^2} \tag{5}$$

where cuisines c_i, $c_j \in C$, the universal set of cuisines in the dataset. To calculate the distance between two cuisines, they must be quantified. This was done by vectorizing the patterns obtained by the above-mentioned pre-processing technique.

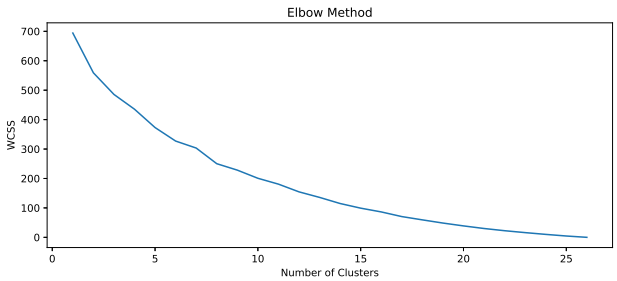

Fig. 1. Elbow Method for cluster identification

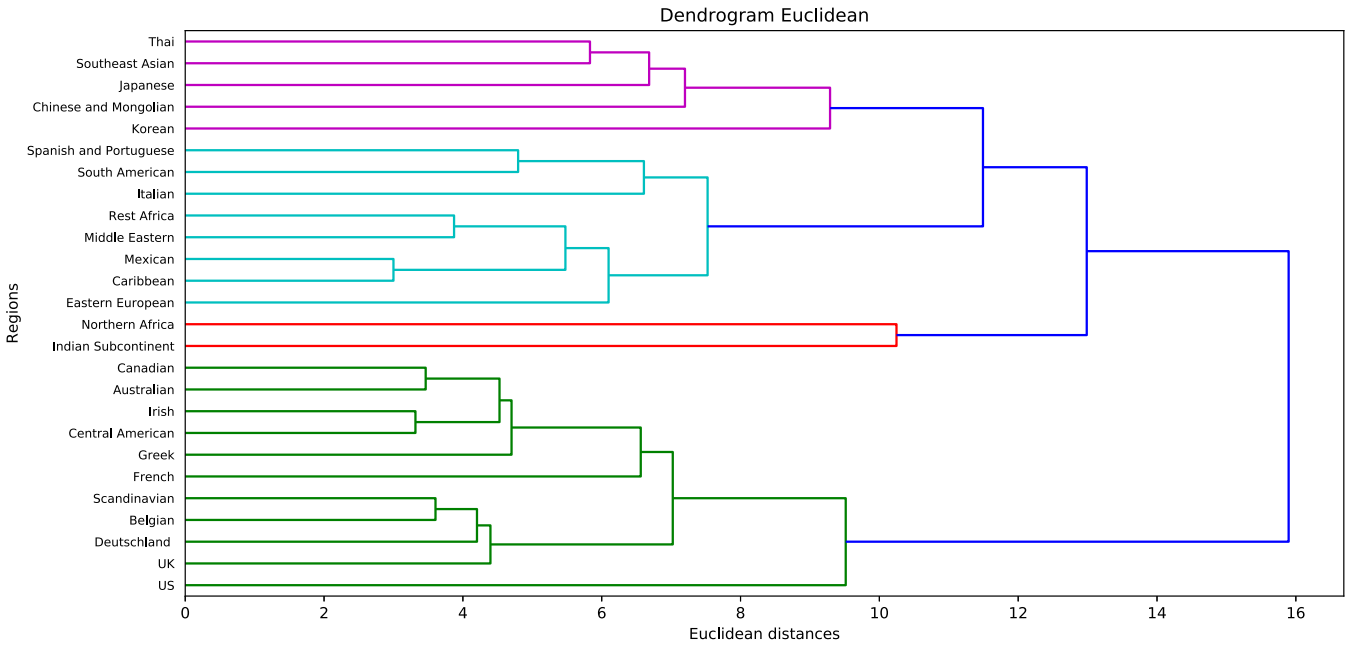

Fig. 2. Hierarchical Agglomerative Clustering based on Patterns Mined using Euclidean distance

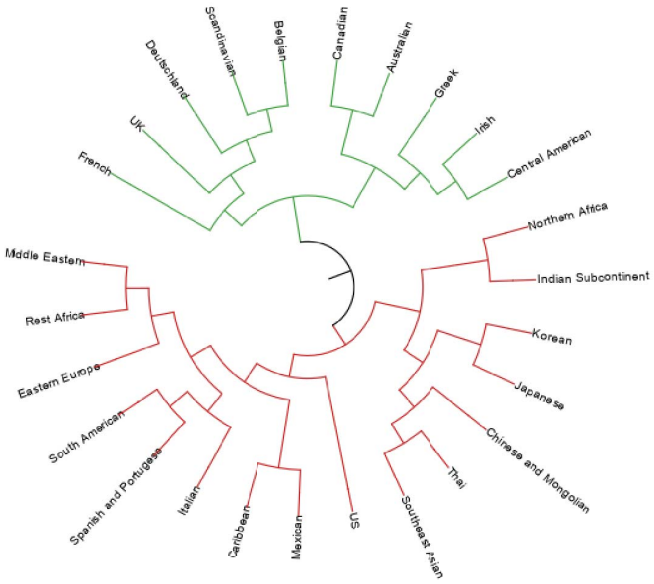

Fig. 3. Hierarchical Agglomerative Clustering based on Patterns Mined using Cosine distance

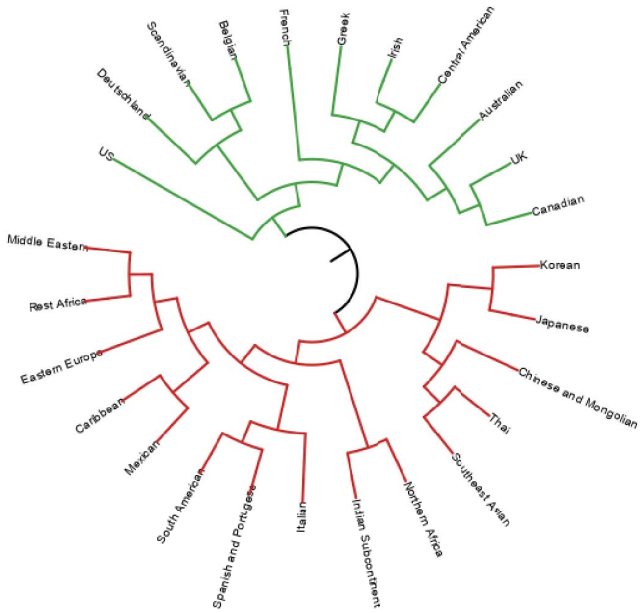

Fig. 4. Hierarchical Agglomerative Clustering based on Patterns Mined using Jaccard distance

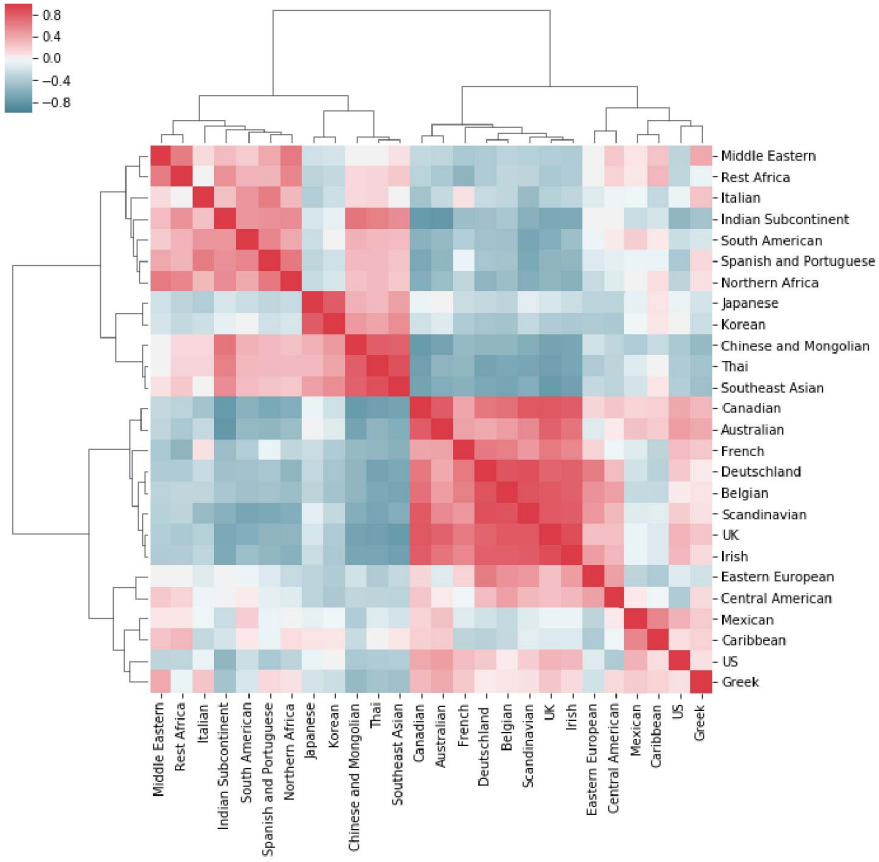

Fig. 5. Hierarchical Agglomerative Clustering based on Authenticity of Ingredients

B. K-means Clustering

Another popular clustering technique, K-means, was applied on our categorical data. It has been shown in [9] that hierarchical agglomerative clustering is a better approach for clustering categorical data than K-means. The elbow analysis and the subsequent WCSS score on our dataset indicates similar results. The elbow method [4] analysis fails to determine the number of appropriate clusters for our dataset. As in Figure 1, no sharp edge or elbow like structure is obtained which determines the number of clusters. While on the other hand the hierarchical agglomerative clustering technique presents with better cluster representation. Therefore, our results were predominantly determined by hierarchical agglomerative clustering.

VII. Results

To evaluate the efficiency of the proposed methodologies, the RecipeDB dataset mentioned in Section III was used. This dataset was used to identify the patterns which were then fed into the Sequential Pattern Mining based clustering algorithm while the ingredients obtained from the dataset were the input features for the Authenticity-based clustering. The corresponding code and relevant files are present in the GitHub repository (https://github.com/cosylabiiit/Hierarchical-Clustering-Ingredients).

The Hierarchical Agglomerative clustering (HAC) gives clusters for all regions based on the three approaches and presents a cluster dendrogram for each approach. Figures 2, 3 and 4 represent the clusters formed using the feature vector obtained via Euclidean, Cosine and Jaccard metrics for pairwise distance calculation respectively. Similarly, Figure 5 shows the authenticity based approach to determine the correlation of cuisine and regions, dominantly based on ingredients.

Because of the absence of a quantified validation metric for cuisine similarity, the geographical relationship among the cuisines was used to validate the accuracy in the prediction of cuisine interrelationships. It is observed that while comparing the Figures 2, 3 and 4 with Figure 6 the results received from the Euclidean distance model were most similar to the geographical distribution of the countries. On the other hand, the clusters obtained via the authenticity based clustering gave similar yet better results than Euclidean distance-based HAC when validated on geographical distance based clusters.

Authenticity-based Clustering identifies both positive and negative relationships between cuisines and items whereas pattern based techniques take only the positive relationships

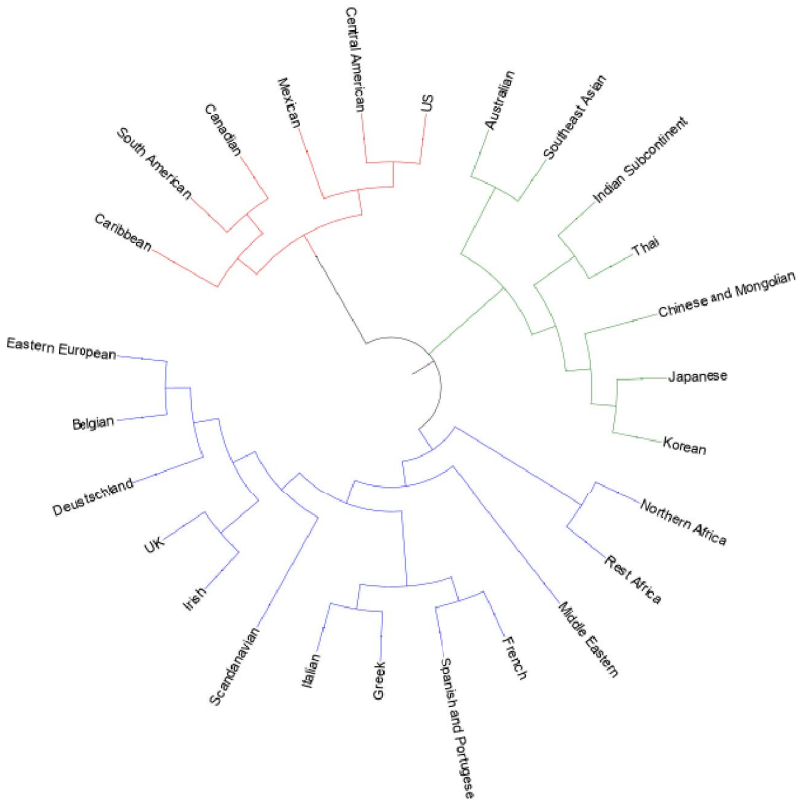

Fig. 6. Hierarchical Agglomerative Clustering based on Geographical Distance of Regions

into account. This leads to the difference in the results. Yet, despite the differences, both techniques predict a closer relationship among Canadian and French cuisines as compared to Canadian and US cuisines despite their geographical proximity. This is evident from the historical fact that Canada was a French colony. Another interesting grouping is that of Indian Subcontinent and Northern Africa. Due to prevalent use of spices in the two regions, Indian subcontinent cuisine is closer to African cuisine as compared to its geographical neighbors like Thai and Southeast Asian cuisines. Hence, the obtained clusters are also able to identify relationships deviating from the geographical similarities.

VIII. Conclusions and Future Work

In this exploratory work we proposed and analyzed two methodologies for fingerprinting cuisines and identifying their interdependence. Our clustering algorithms show how various cuisines are interrelated and show trends similar to their geographical associations. It shows how cooking practices and methods are shared by neighbouring regions. This analysis is important from a historical and cultural point of view as it helps in appreciating how cooking practices are distributed across the world. Furthermore, we also provide a verbose list of patterns identified in the cuisines. These patterns include compound patterns; combination of ingredients, processes and utensils that can be used to identify the relationship among these items.

While this article introduces new methods for investigation of cuisine correlations, it raises new research questions. How do factors such as climate, economy and genetics influence the cuisine patterns? RecipeDB is a sparse dataset in terms of utensils and processes. Hence, to what extent do they influence the relationships among cuisines is yet to be answered. Among one of the limitations of this study, it neither considers the state of ingredients nor their aliases. Therefore, future analysis need to account for the aliases along with state of ingredients and other properties like cooking time and preparation time for the task. It would also be interesting to identify more sophisticated validation metric for cuisine ontology than geographical clustering.

We believe that this study can be applied for cuisine fingerprinting, food ontology and exploration of relations between food and culture. Probing the past and present interrelatedness among cuisines can provide insight into human behavior and

cultures, and means for shaping the future of food.

IX. Acknowledgement

G.B. thanks the Indraprastha Institute of Information Technology (IIIT-Delhi) for providing computational facilities and support. T.S, U.U, J.K,S.A, S.A. and B.A. are Research Interns in Dr. Bagler's lab (Complex Systems Laboratory) at the Center for Computational Biology. All the research interns are thankful to IIIT-Delhi for the support.

References

[1] R. Agrawal, R. Srikant *et al.*, "Fast algorithms for mining association rules," in *Proc. 20th int. conf. very large data bases, VLDB*, vol. 1215, 1994, pp. 487–499.

[2] Y.-Y. Ahn, S. E. Ahnert, J. P. Bagrow, and A.-L. Barabási, "Flavor network and the principles of food pairing," *Scientific reports*, vol. 1, p. 196, 2011.

[3] D. Batra, N. Diwan, U. Upadhyay, J. S. Kalra, T. Sharma, A. K. Sharma, D. Khanna, J. S. Marwah, S. Kalathil, N. Singh *et al.*, "Recipedb: A resource for exploring recipes," *Available at SSRN 3482237*, 2019.

[4] M. J. Brusco and J. D. Cradit, "A variable-selection heuristic for k-means clustering," *Psychometrika*, vol. 66, no. 2, pp. 249–270, 2001.

[5] B. C. Fung, K. Wang, and M. Ester, "Hierarchical document clustering using frequent itemsets," in *Proceedings of the 2003 SIAM international conference on data mining*. SIAM, 2003, pp. 59–70.

[6] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," in *ACM sigmod record*, vol. 29, no. 2. ACM, 2000, pp. 1–12.

[7] S. Hanai, H. Nanba, and A. Nadamoto, "Clustering for closely similar recipes to extract spam recipes in user-generated recipe sites," in *Proceedings of the 17th International Conference on Information Integration and Web-based Applications & Services*. ACM, 2015, p. 31.

[8] A. Jain, N. Rakhi, and G. Bagler, "Analysis of food pairing in regional cuisines of india," *PloS one*, vol. 10, no. 10, p. e0139539, 2015.

[9] A. Joshi and R. Kaur, "A review: Comparative study of various clustering techniques in data mining," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 3, no. 3, pp. 55–57, 2013.

[10] T. Ozaki, X. Gao, and M. Mizutani, "Extraction of characteristic sets of ingredients and cooking actions on cuisine type," in *2017 31st International Conference on Advanced Information Networking and Applications Workshops (WAINA)*. IEEE, 2017, pp. 509–513.

[11] 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*, 2009, pp. 9–14.

[12] N. Singh and G. Bagler, "Data-driven investigations of culinary patterns in traditional recipes across the world," *arXiv preprint arXiv:1803.04343*, 2018.

[13] R. Tuwani, N. Sahoo, N. Singh, and G. Bagler, "Computational models for the evolution of world cuisines," *arXiv preprint arXiv:1904.10138*, 2019.

[14] 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*. ACM, 2011, pp. 105–114.

[15] L. Wang, Q. Li, N. Li, G. Dong, and Y. Yang, "Substructure similarity measurement in chinese recipes," in *Proceedings of the 17th international conference on World Wide Web*. ACM, 2008, pp. 979–988.

[16] S. Yokoi, K. Doman, T. Hirayama, I. Ide, D. Deguchi, and H. Murase, "Typicality analysis of the combination of ingredients in a cooking recipe for assisting the arrangement of ingredients," in *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 2015, pp. 1–6.