New Recipe Recommendation with Visualization

Jiawei Wu, Siya Xie, Ting Wang, Yanxiang Zhou,Yuebai Gao
CSE 6242 Data and Visual Analytics, Project Proposal

Literature survey

Flavor pairing is not only an essence of culinary practice, but also a wisdom. Although rooted in personal experience, the booming of cutting-edge technologies has transformed this ancient subject into an emerging research field, Gatrophysics.[1] The food pairing hypothesis was proposed in 2011 that Western cuisine prefers to use ingredients sharing common flavor components while Asian cuisine avoids.[2] Based on network analysis approaches, this theory opened up an unprecedented field where prediction and construction of successful recipes scientifically is possible under the establishment of pairing rules.[3] Food-bridging was later put forward as an extension to the framework that ingredients can become affine through a chain of food pairs.[3] All these data-driven and network-based approaches were also implemented to other related fields, including nutrition landscapes for balancing nutrient composition in recipe recommendation[4] and deconstructing the region-special cuisine styles.[5,6]

However, most previous research only focused on food-oriented data but overlooked the importance of custom-oriented data source which is also worth mining. Integrating the previous research and both analytic aspects, our group aims at designing a state-of-art website to facilitate recipe exploration and innovation with the application of network algorithms and Machine Learning.

Information networks have been widely used in various disciplines for capturing complex relationships, such as in protein interaction analysis and academic social networks. In protein interaction networks (PINs), centrality analysis is an important task, which is to recognize the most important protein. Jeong et al. proposed the degree centrality method relying on the counts of protein's interactions.[7] However, since edges are unweighted, the results can be affected by false positive interactions. Therefore, a method based on edge clustering coefficient was developed, which indicates tight connection between a protein's neighbors.[8] The PINs were used to predict protein functions, such as identifying key drivers in development of breast carcinomas.[9] These methods inspire us that weighted edges between nodes, which are ingredients in our work, should be more useful for recommendation purpose. To improve upon these studies, we plan to weight the edges by the customer ratings of the recipes that contain both the two ingredients.

Networks are also intensively used to analyze connections and relationships in academia. The networks are formed by entities including publications and scholars, and their relationships, such as citations and co-authorships.[10] Yang et al. introduced a heterogeneous collaboration recommendation network, where multiple features are integrated, including research expertise, co-authorship, and their institutional connectivity.[11] Since research interests of a researcher may change over time, a time-aware topic recommendation network was developed. It was consisted of multiple relationships among users and the temporal information from micro-blogs to profile user's time-drifting topic interests.[12] In addition to collaboration and topic recommendation, a journal recommendation system based on the quality and similarity of manuscripts was also developed.[13] For our work, since recipes also contain multiple levels of entities, like ingredients in a recipe, and nutrients of an ingredient, heterogenous networks could also be useful. Moreover, time-aware analysis could also show the developing and emerging of recipes over time.

Since we have a very large dataset of recipes with hundreds of attributes, one of the most efficient ways to find the connection between the recipes is to use Machine Learning methods on pattern recognition and classification. Pattern recognition and classification have many applications in data mining, including sifting through a large volume of data to extract a small amount of relevant and useful information.[14] However, there is no single learning algorithm that works best on all supervised learning problems, as shown by the no free lunch theorem proposed by Shalev-Shwartz and Ben-David in 2014.[15]

Therefore, in our project, we will try several different classification algorithms, including supervised algorithms like gradient boosting,[16] support vector machine,[17] k-nearest neighbors algorithm and random forest[18]. Studies have shown that all these methods perform well on high dimensional data.

In the data processing part, we will vectorize our recipe data to analyzable feature vectors. If recipe contains the i_{th} ingredient, then its feature vector takes 1 on i_{th} dimension, otherwise 0. By doing this, we can use our model to recognize the pattern of recipes from the perspective of ingredients. Besides, we will also split the original dataset into smaller groups for further analysis.

Research Proposal

1. **What are we trying to do? Motivation and Objective**
   Goals: Our aim is to create a recipe and ingredients visualization tool, providing insights and inspiration for chefs and restaurant runners so that they can use it to create new recipes or get recipes recommendation
2. **How is it done today? And What's new in our approach?**
   Our Approach: Our application will be built on recipe data from food.com. By treating ingredients and other properties as features, we will construct classification and rating prediction models for recipes. Based on such models, we will utilize D3.js to build our interactive web application, conducting recipe analysis, presenting ingredient networks, and giving recipe or ingredient recommendations.
   Related Work and Innovation: Most existing applications relating to the recipe are developed for personal users, and their main job is to recommend existing recipes to users, such as Yummly, Allrecipe.com. While our application is for professional users to create new recipes. On top of that, A launched ingredient recommendation website is Appetit. On the website, given some ingredients, it could generate an ingredient recommendation list based on the food-pairing hypothesis (food shares strong molecular or empirical affinity) (Simas, 2017). However, a similar chemistry compound doesn't guarantee good taste, while our model will be trained over real-world rating data.
3. **Who cares? and What difference and impact will it Make?**
   Our application will provide insights and inspirations for Chefs, independent restaurant owners, amateur cooking lovers and anyone who are interested in creating a new recipe. They can use it to create more delicious food recipes. And we will measure the impact via user studies
4. **What are the risks and payoffs? How much will it cost?**
   Payoffs: more delicious recipe; advertisement revenue if enough user
   Risk: There are 230K recipes in our dataset. A large dataset could lead to a slow loading process and thus affect user experience
   Cost: the primary cost of this project will be the effort of our team. On top of this, a small amount of web hosting cost will be incurred
5. **How long will it take?**
   Approximately 400 hours over 3 months will be invested into this project
6. **What are the midterm and final "exams" to check for success? (Project Timeline)**
   Midterm (November 5th): Data Cleaned, Model constructed
   Final exams (December 3rd): Application Launched, and good feedback from a user study

Activity Plan Table			
Member	Activities	Member	Activities
Jiawei Wu	Proposal*, Front-end coding	Siya Xie	Literature*, UI Design
Ting Wang	Literature*, Front-end coding	Yanxiang Zhou	Literature*, Back-end coding
Yuebai Gao	Presentation*, Front-end coding	Asterisk: done; Unannotated: to be done	

REFERENCES

1 Spence, C. Gastrophysics: Getting creative with pairing flavours. *International Journal of Gastronomy and Food Science*, 100433 (2021).

2 Ahn, Y.-Y., Ahnert, S. E., Bagrow, J. P. & Barabási, A.-L. Flavor network and the principles of food pairing. *Scientific reports* **1**, 1-7 (2011).

3 Simas, T., Ficek, M., Diaz-Guilera, A., Obrador, P. & Rodriguez, P. R. Food-bridging: a new network construction to unveil the principles of cooking. *Frontiers in ICT* **4**, 14 (2017).

4 Kim, S., Sung, J., Foo, M., Jin, Y.-S. & Kim, P.-J. Uncovering the nutritional landscape of food. *PloS one* **10**, e0118697 (2015).

5 Jain, A., NK, R. & Bagler, G. Analysis of food pairing in regional cuisines of India. *PloS one* **10**, e0139539 (2015).

6 Varshney, K. R., Varshney, L. R., Wang, J. & Myers, D. Flavor pairing in Medieval European cuisine: A study in cooking with dirty data. *arXiv preprint arXiv:1307.7982* (2013).

7 Jeong, H., Mason, S. P., Barabási, A.-L. & Oltvai, Z. N. Lethality and centrality in protein networks. *Nature* **411**, 41-42 (2001).

8 Wang, J., Li, M., Wang, H. & Pan, Y. Identification of essential proteins based on edge clustering coefficient. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **9**, 1070-1080 (2011).

9 Hamed, M., Spaniol, C., Zapp, A. & Helms, V. Integrative network-based approach identifies key genetic elements in breast invasive carcinoma. *BMC genomics* **16**, 1-14 (2015).

10 Kong, X., Shi, Y., Yu, S., Liu, J. & Xia, F. Academic social networks: Modeling, analysis, mining and applications. *Journal of Network and Computer Applications* **132**, 86-103 (2019).

11 Yang, C. *et al.* in *2015 48th Hawaii International Conference on System Sciences.* 552-561 (IEEE).

12 Liang, H., Xu, Y., Tjondronegoro, D. & Christen, P. in *Proceedings of the 21st ACM international conference on Information and knowledge management.* 1657-1661.

13 Silva, T., Ma, J., Yang, C. & Liang, H. A profile-boosted research analytics framework to recommend journals for manuscripts. *Journal of the Association for Information Science and Technology* **66**, 180-200 (2015).

14 Dougherty, G. *Pattern recognition and classification: an introduction*. (Springer Science & Business Media, 2012).

15 Shalev-Shwartz, S. & Ben-David, S. *Understanding machine learning: From theory to algorithms*. (Cambridge university press, 2014).

16 Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232 (2001).

17 Cortes, C. & Vapnik, V. Support-vector networks. *Machine learning* **20**, 273-297 (1995).

18 Ho, T. K. in *Proceedings of 3rd international conference on document analysis and recognition.* 278-282 (IEEE).