New Recipe Recommendation with Visualization

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

1. Introduction and Problem Definition

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, Gastrophysics.[1] Several food pairing or bridging systems have been developed, such as Yummly, Food.com, and Appetit. However, most of the existing methods only focused on food-oriented data but overlooked the importance of customer-oriented data source which is also worth mining.

Here, our group aims at designing a state-of-art website to facilitate recipe exploration and innovation. Specifically, we develop a regression machine learning model based on customer ratings of 230,000 + recipes to predict the score of new food combinations, which could guide new recipe recommendation and creation. An interactive food ingredient network will be constructed to visualize the recipe creation process.

Our approach features several innovations. Firstly, the customer rating data, which directly reflects if an ingredient combination is popular or not, is fully utilized to guide the recipe recommendation. Secondly, we use machine learning and regression model to dynamically predict the score of the newly created recipe. More importantly, the food ingredient network realizes dynamic data visualization and facilitate the recipe creation process interactively and intuitively.

2. Literature Survey

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]

Information network is a powerful tool for capturing complex relationships and has been widely used in many disciplines. Centrality analysis is an important task in protein interaction networks (PINs), 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] A method based on edge clustering coefficient was later developed to 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] 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 which featured research expertise, co-authorship, and their institutional connectivity.[11] A time-aware topic recommendation network was developed which consisted of the temporal information from micro-blogs to profile user's time-drifting topic interests.[12] In addition, a journal recommendation system based on the quality and similarity of manuscripts was also developed.[13]

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.

3. Proposed Method

3.1. Data Collection and Cleaning

Our data is a Kaggle dataset crawled from Food.com (GeniusKitchen) online recipe aggregator. Each data record represents a recipe. The attributes we plan to use are tags, ingredients, and ratings. The data cleaning and processing is shown in **Figure 1**. For data cleaning, we mainly used Python to analyze the occurrence of all tags and ingredients, combine same ingredients with different names (e.g., "potatoes" and "new potatoes"), and remove records that have missing information. After data cleaning, we selected several categories of tags, including "time-to-make", "course", "cuisine", and "occasions". For the ingredients, 457 ingredients with a frequency larger than 300 and the recipes containing these ingredients (150,000+ recipes) are kept for further data processing.

3.2. Machine Learning Models

Our back-end model contains two parts: Model A, which is a clustering model based on tags of the recipes, and Model B, which is a group of similar regression models based on ingredient of the recipes. Tags narrow down the range of recipes used for recommendation. Separating tags from other attributes (ingredients) emphasizes the role of tags. And to avoid situations where our recommendations are restricted by certain tags being chosen, we hereby decide not to directly use tags as filters, but instead using the vectorized tags and our Model A to assign a cluster for customers' choice of tags.

Each cluster corresponds to one model in Model B. In Model B, we vectorize the recipes to analyzable feature vectors. The ingredient vectors are one-hot encoded. Since 457 ingredients are kept, each recipe is turned into a vector with 457 dimensions. Each ingredient corresponds to one dimension of the recipe vector. If a recipe contains the i_{th} ingredient, then its feature vector takes 1 on i_{th} dimension, otherwise 0. These vectors along with the customer rating data are used for the regression model learning and rating score prediction. The specific features of Model A and Model B are listed in the following table.

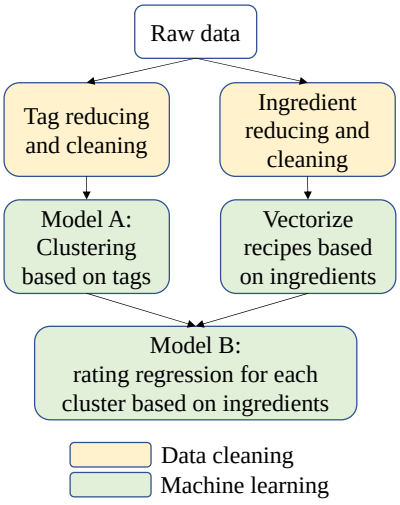

Figure 1. Flow chart of data cleaning and processing

	MODEL A	MODEL B
Type	Clustering	Regression
Input data	Tags choose by user	Ingredients choose by user/cluster ID
Output data	Cluster ID	Rating prediction
Candidate methods	K-means, DBSCAN, Gaussian mixture model, Spectral clustering	xgboost, SVM, Random forest, Ridge regression, LASSO, Linear regression
Details	1. Vectorize the input tags. 2. Use the model to identify which cluster the vector belongs to. 3. Assign a cluster id.	1. Use cluster id to find the corresponding model. 2. Add only one more ingredient from ingredient list each time. 3. Vectorize the modified input data. 4. Feed ingredient vector to the selected model. 5. get rating prediction, find top k ingredients yields largest rating prediction

3.3. User Interface Design and Website Construction

Hompage

The homepage consists of four sections, welcome, ingredient list, recipe tags, and start. The first section (**Figure 2**) introduces our website to users, including the coridea of our project and primary coding tools applied, together with our affiliation and website usage.

The name of our website "CooNet" stands for a combination of "Cool" and "Cook".

Scrolling down then comes the page for users to input their own ingredient list, based on which new ingredients are added through computation. Users can search and browse the provided ingredient list on the left and add them into or remove chosen ingredients from their user-defined list on the right.

After the ingredient list is the page for recipe tags. We classified all recipes in our database into several dominant genres and provide users tags for each category. Checking tags fitting to their expected recipe, prediction process is optimized by limiting the database for training.

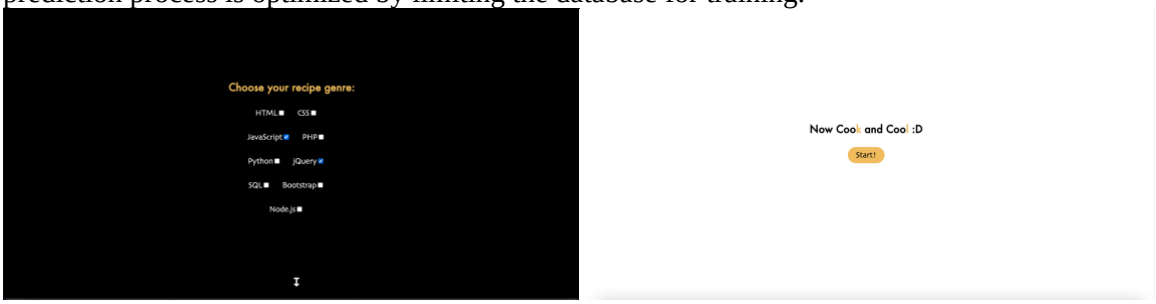

Figure 2. Homepage designs. All sub-figures shown here are screenshots of webpages already coded and compiled. Future replenishment and modification are required. (upper left: Intro, upper right: ingredient list, bottom left: recipe tags, bottom right: start)

Result page

After clicking the "start" button in the last section, our platform computes the result based on the input defined by users in the homepage and then generates the result page. The left large canvas exhibits the ingredient network (**Figure 3**) composed of user-defined nodes (black) and algorithm-predicted nodes (white). By clicking on the algorithm nodes, users can add them into the user-defined ingredient list and thus replenish their recipe.

Statistics of recipes satisfying the same styling condition (tags) are visualized in the inset figures placed at the top-right corner. Users can switch the variable of the histogram and also slide the scale thumb to filter recipes listed below.

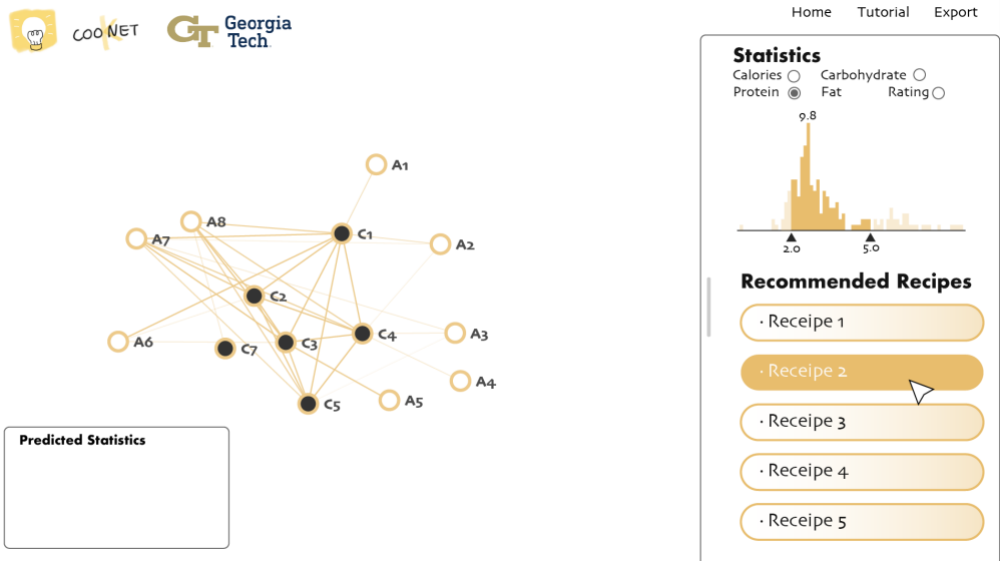

Figure 3. Result page design. Figure shown here is a design sketch of the future result page made in Adobe Illustrator. There will be discrepancy between the sketch and final design.

3.4. Connection between frontend visualization with backend data processing

The controller of the web application will be implemented using the python application, Flask. After the frontend requests certain data, the controller is supposed to organize the required data into json format before sending it to frontend JavaScript D3. To publish the website for more convenient delivery, we plan to host the website on AWS.

4. Design of Experiments and Evaluation

The evaluation will be centered on whether the product accomplished the project's objectives as stated in the introduction. Hence, the main questions will be answered by experiments are:

- How accurate are the rating prediction made by the regression algorithm?
- Is the visualization user friendly?

Evaluation of prediction algorithm

To determine the accuracy of our algorithm, we will randomly divide the recipe dataset into two parts. 75% of recipes will be used as training dataset and 25% as test dataset. The overall accuracy will be measure by Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{\frac{\sum_{i=1}^{test} \left(p^{(i)} - y_{test}^{(i)}\right)^2}{N_{test}}} \qquad (1)$$

Evaluation of User Experience

We will design a user experience survey to find out how well our product meets user's need. The sample size will be set to 40 users. And some question in the survey will include:

- Are you satisfied with the interface and visualization of the website?
- Do you like the ingredients recommended by our product?
- Do think it useful when you are creating a new recipe?

5. Plan of Activities

The main work includes data collection and cleaning, clustering and regression model building, UI design and data visualization coding. We plan to finish data cleaning by 11/10, build up the machine learning modeling by 11/17, construct the website by 11/28, and finish evaluation, report writing by 12/3. The activity distribution among team members is listed in the following table.

Member	Activities
Jiawei Wu	Proposal*, UI Design*, Front-end UI coding [JavaScript]
Ting Wang	Literature*, Back-end data cleaning* & processing [Python]
Yuebai Gao	Presentation*, Front-end UI coding [JavaScript], controller coding [Python]
Siya Xie	Literature*, UI Design*, Front-end UI coding [JavaScript]
Yanxiang Zhou	Literature*, Back-end data cleaning* & processing [Python]

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).