Postprint: Research on Mining Dietary Communities Based on Recipes and Weibo User Comments

Authors: Wu Xiaolan, Zhang Chengzhi

Date: 2017-10-11T00:00:00+00:00

Abstract

【Objective】 This study investigates dietary community structure supported by large-scale real social network data. 【Method】 Using recipe information from the "Meishijie" website and dish-related Weibo data from Sina Weibo, we constructed "mention" relationships between users and dishes, mapped them at both provincial/regional and regional cuisine dimensions, and applied community detection algorithms for community mining. 【Results】 Distinct community structures exist in both the provincial/regional relationship network and the regional cuisine relationship network. 【Limitations】 During the experiment, the significant disparity in population numbers between developed and marginal regions has a certain impact on the conclusions drawn in this paper. 【Conclusion】 Empirical results reveal that provinces are divided into three flavor regions: "Other Flavors," "Fresh-Salty," and "Spicy-Fragrant" ; "Sichuan Cuisine" and "Yunnan-Guizhou Cuisine" are rarely ordered together with other cuisines due to their unique auxiliary ingredients, while "Beijing Cuisine," "Shanghai Cuisine," "Shandong Cuisine," and "Northeastern Cuisine" are frequently ordered together; additionally, there exists a certain degree of geographic proximity among regional cuisines.

Full Text

Analyzing Food Community with Recipes and Weibo User Reviews

Wu Xiaolan[1],[2], **Zhang Chengzhi**[2],[3]
[1](School of Management Science and Engineering, Anhui University of Finance and Economics, Bengbu 233030, China)
[2](Department of Information Management, Nanjing University of Science and Technology, Nanjing 210094, China)

[3](Jiangsu Key Laboratory of Data Engineering and Knowledge Service, Nanjing 210093, China)

Abstract

Objective: This study examines the structure of online food communities supported by large-scale real-world social network data. **Methods:** We utilized recipe information from the "Meishijie"website and dish-related Weibo data from Sina Weibo. After constructing "mention"relationships between users and dishes, we mapped these relationships across provincial regions and regional cuisine dimensions, then applied community detection algorithms for mining. **Results:** Clear community structures emerged in both the provincial region network and the regional cuisine network. **Limitations:** The significant disparity between user populations in developed versus peripheral regions during the experiment may influence our conclusions. **Conclusions:** Empirical findings reveal that provincial regions cluster into three taste zones: "Other Flavors," "Fresh-Salty Flavors," and "Hot-Spicy Flavors." "Sichuan Cuisine" and "Yungui Cuisine" are rarely ordered together with other cuisines due to their unique ingredients, while "Jing Cuisine," "Hu Cuisine," "Lu Cuisine," and "Northeastern Cuisine" are frequently ordered together. Additionally, regional cuisines exhibit a certain degree of geographical proximity.

Keywords: Food culture; Regional cuisines; Food community; Web information organization
Classification Number: G353

Food is an eternal theme of human society and the foundation for other social activities. With the development of productivity, especially agriculture, Chinese people' s relationship with "eating" has evolved far beyond mere sustenance, giving rise to a comprehensive food culture encompassing customs, philosophies, and behaviors. As a cultural phenomenon related to "eating" and "drinking," food culture transcends race and nationality and concerns everyone. Therefore, research on Chinese food culture is both necessary and valuable.

Although current food culture research spans various disciplines with diverse methodologies and abundant findings, most studies employ qualitative analysis of recipes (such as tracing the evolution of food culture) and quantitative analysis (such as statistical compilation of food culture literature and historical records). In reality, with the development of Internet and big data technologies, large-scale real-world datasets enable food-related research at unprecedented scales, allowing verification of important conclusions and discovery of new, practically valuable insights through data mining. This paper therefore investigates food community mining based on recipe data and Weibo user dietary comments.

2. Related Research Overview

Currently, food culture research in China is primarily conducted by scholars from higher education institutions, food industry practitioners, and literary writers. Compared to other disciplines, academic research on food culture started relatively late, and only in recent years have scholars begun utilizing real-world data. Summarizing existing studies, we identify two main types of real-world data commonly used: recipe data and user dietary comment data.

Corresponding Author: Zhang Chengzhi, ORCID: 0000-0001-8121-4796, E-mail: zhangcz@njust.edu.cn.

This work is supported by the National Social Science Foundation Project "Research on User-Based Knowledge Organization Models in Online Social Networks" (Project No.: 14BTQ033), the Anhui Provincial Department of Education Humanities and Social Science Project "Research on Interdisciplinary Knowledge Discovery and Its Applications Based on Social Networks" (Project No.: SK2016A0025), and the Jiangsu Key Laboratory of Data Engineering and Knowledge Service Open Project "Research on Interdisciplinary User Knowledge Structure Discovery and Interest Evolution in Online Social Networks" (Project No.: DEKS2014KT006).

Among these two data types, studies utilizing recipe data include: Wagner et al. constructed a flavor network through co-occurrence relationships of cooking ingredients in common compound flavorings, discovering that Western cooking tends to use multiple spices to create mixed flavors, satisfying the food pairing hypothesis, while East Asian cooking does the opposite. Ahn et al. analyzed 56,498 recipes from multiple countries and regions, revealing significant differences between Western and Eastern diets—the six most popular ingredients in the West are milk, butter, vanilla, eggs, sugar syrup, and wheat, while in the East they are soy sauce, scallions, sesame oil, rice, soybeans, and ginger. Moreover, Western chefs prefer combining ingredients with many shared spices, whereas Eastern chefs avoid this practice. Zhu et al. collected 8,498 recipes and 2,911 ingredients from 20 cuisines on the Meishijie website, combined with geographical location and climate similarity of cuisine origins, finding that geographical proximity influences ingredient usage far more than climate similarity. Additionally, using simple principal component analysis on the ingredient usage matrix (with cuisines and ingredients as dimensions), they identified Yungui Cuisine and Hong Kong Cuisine as anomalous cuisines.

Studies utilizing user dietary data include: Ahn et al. analyzed user dietary preferences on a large recipe website (ichkoche.at) from August 2012 to November 2013, finding that user recipe preferences primarily depend on dish spices, that regional differences in recipe preferences exceed those in spice preferences, and that weekday dietary preferences differ significantly from weekend preferences. Abbar et al. analyzed dietary tweets from 210,000 Twitter users along with user interests, geographical locations, and social network data, discovering a correlation between food calories and local obesity rates with a Pearson corre-

lation coefficient close to 0.77, and built models to predict regional obesity and diabetes rates based on demographic variables and food names mentioned on Twitter.

In summary, we find that research on food community mining remains scarce. Food community mining can not only deeply explore regional user tastes but also discover user ordering patterns, providing guidance for meal selection. Therefore, this paper combines recipe data with user dietary comment data to investigate food community mining.

3. Research Approach and Key Technologies

3.1 Research Approach

This study investigates food communities by combining recipe data and Weibo user dietary comments. The research framework is illustrated in Figure 1 [Figure 1: see original paper].

Figure 1. Research Framework for Food Community Mining

(1) **Data Collection and Preprocessing.** We collected recipe names, cuisine categories, and other information from the Meishijie website. After simple preprocessing of the dish names, we used them as search keywords to crawl Weibo content and user information from Sina Weibo. Following user data collection, we added the recipe names to the segmentation dictionary and performed effective word segmentation on user comment content.

(2) **Relationship Mapping.** After segmenting user comments, we extracted "mention"relationships between users and dish names (where a recipe name appears in a user' s Weibo comment). Based on user provinces and dish cuisines, we performed two mappings:

- **Province-Province Relationship Mapping:** Since users have provincial information, if users from different provinces both "mention" the same dish, we can infer a common connection between these provinces, such as similar user tastes. Therefore, we mapped province-province relationships based on user-dish "mention" relationships.
- **Cuisine-Cuisine Relationship Mapping:** Similarly, since dishes have cuisine information, if the same user "mentions" different cuisines, these cuisines likely share some correlation. Thus, we mapped cuisine-cuisine relationships based on user-dish "mention" relationships.

(3) **Community Discovery.** After completing province-province and cuisine-cuisine relationship mapping, we selected appropriate community mining algorithms to discover food communities and analyze implicit relationships between provinces and between cuisines, followed by results visualization.

3.2 Key Technologies

Community discovery is the core technology in this research. Numerous community detection algorithms exist, including classic methods such as Girvan and Newman' s GN divisive algorithm, Newman' s modularity maximization method, Shi's normalized cut (N-cut) method, Von Luxburg's spectral partitioning method based on Laplacian matrices, and the LPA algorithm. The Label Propagation Algorithm (LPA), proposed by Zhu et al. in 2002 as a graph-based semi-supervised learning method, uses labeled node information to predict unlabeled nodes. In 2007, Raghavan et al. first applied LPA to community detection, testing it on real benchmark networks including Zachary' s Karate network and the College Football network, demonstrating its effectiveness.

The LPA steps for community detection are: 1. Initialize labels for all network nodes, assigning each node a unique label. 2. Set iteration count t=1. 3. Randomly permute network nodes to generate sequence X. 4. For each node v in sequence X, update its label using arg max |N (v)|, where N (v) is the set of neighbor nodes of v with label l. If multiple labels have the maximum count, randomly select one. 5. If every node' s label is the most frequent among its neighbors, the algorithm stops; otherwise, set t=t+1 and return to step 3.

For community partitioning, modularity proposed by Newman et al. in 2004 measures partition quality. The modularity formula is:

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(C_i, C_j)$$

where A_{ij} is the adjacency matrix, m is the total number of edges, $P_{ij} = \frac{k_i k_j}{2m}$ represents the expected number of edges between vertices i and j in a null model, and $\delta(C_i, C_j) = 1$ if vertices i and j are in the same community, otherwise 0. This formula can also be applied to weighted graphs after appropriate calculations on the adjacency matrix and node degrees, which we use as our community partitioning criterion.

4. Experiments

4.1 Experimental Dataset

We collected recipes from 20 cuisines on the domestic website "Meishijie" [1], including dish names, categories, cuisines, main ingredients, auxiliary ingredients, and cooking methods. After filtering dish names from Zhu et al.[2], we obtained 5,156 valid recipes across 20 cuisines: Sichuan, Northeastern, Hong Kong-Taiwan, Other, Hubei, Shanghai, Anhui, Jiangxi, Beijing, Shandong, Fujian, Halal, Shanxi, Jiangsu, Northwestern, Hunan, Henan, Cantonese, Yungui, and Zhejiang. The distribution is shown in Table 1 .

Table 1. Recipe Count Statistics by Cuisine

Using these dish names as search queries on Weibo, we collected 8,746,931 Weibo posts from 3,980,597 users across 36 different regions (including "Other" and "Overseas" as user-reported locations). Each Weibo post included publication time, content, user province, and gender. After adding preserved dish names to the Jieba segmentation dictionary and segmenting user dietary Weibo content, we extracted dish names from the segmentation results, obtaining 2,269,763 user-dish "mention" relationships as our primary research object for community discovery. Table 2 shows the user count statistics by region.

4.2 Relationship Mapping

To analyze provincial user tastes and cuisine ordering patterns, we mapped these "mention" relationships across regional and cuisine dimensions, obtaining region-region weighted graphs and cuisine-cuisine weighted graphs. Despite the large scale of user-dish "mention" relationships, the resulting weighted graphs were limited in size but had substantial edge weights. Tables 3 and 4 show the closest nodes (maximum weight) for each region and cuisine.

Table 2. User Count Statistics by Region

Table 3. Closest Regions and Weights in Region Mapping Graph

Table 4. Closest Cuisines and Weights in Cuisine Mapping Graph

Table 3 reveals that, excluding the unknown regions "Other" and "Overseas," Zhejiang, Jiangsu, Guangdong, and Beijing show extremely strong connections with other provinces, likely due to larger user populations in these areas. Further analysis of cuisine closeness in Table 4 shows that among the 20 cuisines, 18 have the highest co-occurrence frequency with themselves, while only "Jing Cuisine" and "Sichuan Cuisine"show the highest co-occurrence with each other, indicating Sichuan Cuisine' s popularity.

4.3 Food Community Mining Results

After mapping, we attempted community detection on the complete weighted region and cuisine graphs. However, we found these complete weighted graphs could not be partitioned into communities, as modularity remained 0 after multiple iterations. This likely occurs because complete weighted graphs lack typical community structures where intra-community connections are relatively dense while inter-community connections are relatively sparse. Therefore, we first performed edge-cutting on these complete weighted graphs to disconnect some node connections. Edge-cutting is commonly used in network virus propagation control to effectively inhibit spread. We applied it here because our complete graphs were unsuitable for community partitioning.

Based on our analysis, we observed that our complete graphs have edge weights (reflecting co-occurrence counts between different nodes) and node weights (reflecting self co-occurrence counts). According to Wu et al., only nodes with approximately equal weight values can exhibit synchronization or behavioral

correlation. Thus, if node weight differences are substantial, we argue these nodes cannot be partitioned into the same community (which typically consists of functionally similar or property-similar network nodes). We therefore cut edges between nodes with 悬殊 weight differences, using the criterion in formula (2):

$$|W_p - W_q| > W_{Epq}$$

where W_p and W_q represent the weights of nodes p and q, and W_{Epq} is the weight of edge E_{pq} between them.

After edge-cutting, the region complete weighted graph with 666 edges (including self-connections) and the cuisine complete weighted graph with 210 edges were reduced to 261 and 40 edges, respectively. We then applied LPA to the edge-cut weighted graphs, executing 100 runs and selecting results with maximum modularity: Figure 2 Figure 2: see original paper ($Q = 0.370$) for regions and Figure 3 Figure 3: see original paper ($Q = 0.695$) for cuisines. Pre-partitioning results are shown in Figures 2(b) and 3(b), where node sizes reflect self co-occurrence weights and edge thickness reflects inter-node co-occurrence weights.

Figure 2. Region-Region Weighted Network Graph

Figure 3. Cuisine-Cuisine Weighted Network Graph

Figure 2 shows three communities (nodes with same color belong to one community): 1. **Community 1:** Qinghai, Macau, Ningxia, Tibet 2. **Community 2:** Guizhou, Guangxi, Yunnan, Gansu, Jiangxi, Hainan, Tianjin, Hebei, Jilin, Heilongjiang, Hunan, Taiwan, Inner Mongolia, Anhui, Shaanxi, Shanxi, Xinjiang, Hong Kong 3. **Community 3:** Liaoning, Beijing, Guangdong, Shanghai, Shandong, Fujian, Other, Overseas, Sichuan, Chongqing, Hubei, Jiangsu, Henan, Zhejiang

The partitioning reveals that: 1. Community 1 consists mainly of autonomous regions 2. Community 3 comprises China' s more developed regions (Beijing, Shanghai, Guangdong, Jiangsu, Zhejiang, Fujian), suggesting these areas discuss similar dishes (i.e., share similar tastes)

To further compare tastes within each community, we 统计了 the most frequently mentioned dishes and their mention counts across provinces in each community, shown in the left side of Table 5 . We also extracted distinctive regional dishes through set difference operations on the top 100 dish names within each community, with results on the right side of Table 5.

Table 5. Statistics of Dishes Mentioned by Provincial Users

The left data shows common preferences across communities for dishes like "grilled meat," "sour and spicy noodles," and "maoxuewang" (a Sichuan dish). The right data reveals distinctive taste differences: Community 1' s unique

dishes belong to "Other Flavors," Community 2' s to "Fresh-Salty Flavors," and Community 3' s to "Hot-Spicy Flavors."

Figure 3 shows that cuisines partition into nine communities based on co-occurrence relationships: 1. **Community 1:** Hunan, Hubei, Northwestern, Halal, Zhejiang 2. **Community 2:** Beijing, Shanghai, Shandong, Northeastern 3. **Community 3:** Cantonese, Anhui, Jiangsu 4. **Community 4:** Shanxi, Henan 5. **Community 5:** Hong Kong-Taiwan, Jiangxi 6. **Community 6:** Fujian 7. **Community 7:** Sichuan 8. **Community 8:** Yungui 9. **Community 9:** Other

The results show that: 1. "Sichuan Cuisine," "Fujian Cuisine," "Yungui Cuisine," and "Other Cuisine" form independent communities. The isolation of "Yungui Cuisine" and "Other Cuisine" aligns with findings in literature [4]. 2. Communities 1, 3, and 4 demonstrate geographical proximity among cuisines. According to Zhu et al., "Northwestern Cuisine" originates from Shaanxi, Gansu, Qinghai, and Ningxia; "Halal Cuisine" from Xinjiang; and "Henan Cuisine" from Shandong—confirming geographical proximity.

To further analyze why "Sichuan Cuisine" and "Yungui Cuisine" form independent communities and to explain Communities 2 and 5, we analyzed auxiliary ingredient proportions at the ingredient level (calculated as total auxiliary ingredients divided by number of dishes per cuisine), with results in Table 6 .

Table 6. Statistics of Auxiliary Ingredient Usage Ratios by Cuisine

Table 6 reveals that "Sichuan Cuisine"is indeed "spicy"(frequently using "Sichuan peppercorn," "pepper," "chili," and "doubanjiang" beyond common condiments), while "Yungui Cuisine" is distinctive for rarely using "MSG" but frequently using "shrimp" —justifying their independent community status. Community 2' s "Beijing," "Shanghai," "Shandong," and "Northeastern" cuisines show similar tastes, particularly "Northeastern Cuisine" as a branch of "Shandong Cuisine." Community 5' s grouping of "Hong Kong-Taiwan" and "Jiangxi" cuisines is difficult to explain through ingredients, though "Hong Kong-Taiwan Cuisine" prominently features "stock" as a main auxiliary ingredient.

This study provides preliminary provincial and cuisine partitions without deeply mining the underlying associations within each community. Additionally, we only considered "mention" relationships without analyzing sentiment polarity, which represents a direction for future research.

References

[1] Chen Guolin. Study of Food Culture: An Overview and Its Constraints [J]. Journal of Sichuan Higher Institute of Cuisine, 2013(2): 4-7.

[2] Wagner C, Singer P, Strohmaier M. The Nature and Evolution of Online Food Preferences [J]. EPJ Data Science, 2014, 3(1): Article No. 38.

[3] Ahn Y Y, Ahnert S. The Flavor Network [J]. Leonardo, 2013, 46(3): 272-273.

[4] Zhu Y X, Huang J, Zhang Z K, et al. Geography and Similarity of Regional Cuisines in China [J]. PLoS ONE, 2013, 8(11): e79161.

[5] Ahn Y Y, Ahnert S E, Bagrow J P, et al. Flavor Network and the Principles of Food Pairing [OL]. arXiv: 1111.6074.

[6] Abbar S, Mejova Y, Weber I. You Tweet What You Eat: Studying Food Consumption Through Twitter [OL]. arXiv Preprint, 2014. arXiv: 14124361.

[7] Girvan M, Newman M E. Community Structure in Social and Biological Networks [J]. Proceedings of the National Academy of Sciences, 2002, 99(12): 7821-7826.

[8] Newman M E, Girvan M. Finding and Evaluating Community Structure in Networks [J]. Physical Review E, 2004, 69(2): 026113.

[9] Shi J, Malik J. Normalized Cuts and Image Segmentation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(8): 888-905.

[10] Von Luxburg U. A Tutorial on Spectral Clustering [J]. Statistics and Computing, 2007, 17(4): 395-416.

[11] Raghavan U N, Albert R, Kumara S. Near Linear Time Algorithm to Detect Community Structures in Large-scale Networks [J]. Physical Review E, 2007, 76(3): 036106.

[12] Zhu X, Ghahramani Z. Learning from Labeled and Unlabeled Data with Label Propagation [R]. Carnegie Mellon University, 2002. http://discovery.ucl.ac.uk/id/eprint/185718.

[13] Zachary W W. An Information Flow Model for Conflict and Fission in Small Groups [J]. Journal of Anthropological Research, 1977, 33(4): 452-473.

[14] Newman M E. Analysis of Weighted Networks [J]. Physical Review E, 2004, 70(5): 056131.

[15] Song Yurong, Jiang Guoping, Xu Jiagang. An Epidemic Spreading Model in Adaptive Networks Based on Cellular Automata [J]. Acta Physica Sinica, 2011, 60(12): 110-119.

[16] Wu Liang, Zhu Shiqun. Node Weights and Its Physical Significance in Networks [C]. In: Proceedings of the 12th National Symposium on Quantum Optics. 2006.

Author Contributions

Wu Xiaolan: Literature research and organization, experimental implementation, draft writing.
Zhang Chengzhi: Research conceptualization, discussion of research methodology, revision of final manuscript.

Conflict of Interest Statement

All authors declare no conflict of interest.

Supporting Data

Supporting data is self-archived by the authors, E-mail: wuxiaolananhui@163.com.

[1] Wu Xiaolan, Zhang Chengzhi. userid&province.txt. User IDs and their mentioned dish names and cuisines.
[2] Wu Xiaolan, Zhang Chengzhi. userid&dietname.txt. User IDs and user provinces.
[3] Wu Xiaolan, Zhang Chengzhi. province_{network}.txt. Province network after regional dimension mapping.
[4] Wu Xiaolan, Zhang Chengzhi. diet_{network}.txt. Cuisine network after cuisine dimension mapping.
[5] Wu Xiaolan, Zhang Chengzhi. diet_{network}(cut_{edge}).txt. Cuisine network after edge-cutting.
[6] Wu Xiaolan, Zhang Chengzhi. province_{network}(cut_{edge}).txt. Province network after edge-cutting.

Received: 2016-03-17
Revised: 2016-03-30

CCC Enhances RightFind Content Workflow Solution

Copyright Clearance Center, Inc. (CCC), a company dedicated to creating global licensing and copyright content solutions, recently announced enhancements to its cloud-based RightFind content workflow solution.

RightFind provides users with instant, convenient access to thousands of journals while helping administrators optimize procurement and management expenditures. RightFind 7.0 includes three main functional enhancements: 1. An upgraded user interface to streamline workflows and make content discovery easier 2. Integration of CrossRef data to accelerate the revelation of citation information for recently published literature in RightFind 3. Addition of two new APIs allowing users to extract information from the RightFind repository for use in other applications to search RightFind content

Lauren Tulloch, CCC' s Director of Products and Services, stated: "Our customers seek seamless connections with information. We are strengthening our platform to enhance the RightFind user experience while making it easier for other applications to extract and utilize data from RightFind."

As part of the RightFind content workflow solution components, CCC also provides RightFind XML for text mining. Researchers in life sciences can create full-

text XML files from over 5 million articles across more than 6,000 peer-reviewed journals for copyright-compliant use in third-party text mining software.

(Compiled from: http://www.copyright.com/copyright-clearance-center-announces-latest-enhancements-to-rightfind-content-workflow-solution/) (Journal News)

Note: Figure translations are in progress. See original paper for figures.