Cluster analysis of Xi'an restaurants by Self-organizing maps

Jie Kong, Meng Ren
School of Computer Science, Xi’an Shiyou University, Xi’an, Shaanxi, 710065, China
jkong@mail.xsyu.edu.cn

Abstract. To investigate the segmentation features of restaurants in Xi’an, China, the Self-organizing maps are applied in this study to analyse the information of online reviews obtained from Dazhong Dianping, which is a famous Chinese social network. Through the clustering and visualization performed by Self-organizing maps, 10 segments of restaurants are identified and some representative features are summarized. The findings of this study could help managers to improve restaurant competitiveness, as well as customers’ decision making.

1. Introduction
With the rapid development of Internet, lots of new web applications have emerged in the past decade. One of the most commonly applied web applications in the daily life of Chinese people is the online local life information service, which provides the information of stores, customer reviews, discount coupons, etc. On the other hand, the information published on the online local life information platform, such as the customer reviews, directly impacts the decision making of potential customers and the business performance of the related stores. Therefore, how to analyse this online information accurately and efficiently becomes a challenge for catering industry practitioners.

In recent years, various data mining techniques have been widely applied in business analysis to provide insightful knowledge from the data generated by commercial activities. Among these techniques, clustering is an important and popular approach in tourism market segmentation, which could divide the collection of travellers, attractions, services, stores, etc. into multiple segments consisting of similar features [1]. It provides better understanding of tourism products for tourism managers to enhance their business competitiveness, as well as for travellers and customers to assist their decision making. To investigate the segmentation features of restaurants located in Xi’an, China, Dazhong Dianping, a famous Chinese online local life information website for customer reviews of restaurants and entertainment, is adopted as the data source for the cluster analysis. Restaurants information and ratings generated by customers are collected by web spider and then analysed by a clustering algorithm named Self-organizing maps (SOM).

2. Related Works
At present, many data mining techniques have been applied to analyse the online reviews of tourism products and services. For instance, Support Vector Machine (SVM) is applied to identify what kinds of user generated messages of restaurants on Facebook are the most popular reviews, which obtain the largest number of comments and “Like”s [2]. Geotagged photos are collected from online travel diaries and analysed by sequential rule mining to provide deep insights for tourists’ preference [3]. Travellers’ review sentiments are classified by applying Naive Bayes on the online reviews of three major tourism and hospitality review platform: TripAdvisor, Expedia, and Yelp [4]. Artificial Neural Networks (ANN) are adopted to identify the key determinants of hotel performance through online reviews [5]. Hotel
customer expectations are analysed by applying association rules on customer reviews of social networks [6]. Geotagged photos posted on social networks are analysed by DBSCAN algorithm to identify tourism attractions [7]. To identify customers’ emerging hotel preference, text mining is used to analyse customer online reviews [8].

As an important issue in tourism and hospitality, investigating the online reviews of restaurants through data mining approaches have also attracted the attention of academia. The existed researches in this stream mainly concentrate on customer sentiment identification. For example, customers’ sentiments of restaurants are identified from online reviews through supervised learning algorithms such as Naïve Bayes, SVM [2, 9], by text mining approach such as calculating TFIDF of words in online reviews [10] or data mining tools [11], through unsupervised learning approach such as CHAID [12] and k-means [13]. However, to the best of our knowledge, there are only a few researches work on the market segmentation problems through the analysis of restaurant online reviews with the assists of data mining. Liao et al. try to infer the styles of restaurants from online photos posted on social network by Convolutional Neural Network (CNN) [14]. This study does not provide a comprehensive solution for the segmentation of restaurant information. In view of this, SOM is applied in this study to investigate the online reviews of restaurants in a famous Chinese online local life information platform.

3. Methodologies

3.1. The collection of data
To collect online reviews for data analysis, a famous Chinese online local life information website named Dazhong Dianping (www.dianping.com) is selected as the data source in this study. This website allows customers to publish five-level ratings on overall, tastes, environment and services, as well as comments and photos. Xi’an, an ancient Chinese city which is famous for tourism, culture, history and food, is selected as the research object of this study. The information of restaurants in Xi’an published on Dazhong Dianping are collected by a web crawler program, which is developed by Python and based on a web crawling framework named Scrapy [15]. Attributes of the restaurant information collected by the crawler program is shown in table 1. At last, a total of 10908 data records are obtained by the crawler program and used for data analysis.

| Attributes          | Comments                                           |
|---------------------|----------------------------------------------------|
| Star                | The overall rating of a restaurant                 |
| taste_grade         | The rating of taste of a restaurant                |
| surrounding_grade   | The rating of environment of a restaurant          |
| service_grade       | The rating of service of a restaurant              |
| address             | The location of a restaurant                       |
| type                | The category of cuisine of a restaurant            |
| consume             | Per capita consumption in a restaurant            |
| evaluate_num        | The number of reviews of a restaurant              |

3.2. The basis of Self-organizing maps
The collected information contains several attributes of restaurants, how to identify the segmentation feature of these restaurants? In this study, the SOM algorithm is adopted to do the clustering and visualization of restaurants information. The SOM is an unsupervised learning algorithm which is usually used to map the high dimensional data into low dimensional space while keeping the topological structure of data unchanged [16]. Therefore, it is suitable for dimension reduction, visualization and clustering. A typical SOM is consisting of two layers: the Kohonen layer and the input layer. As shown in Figure 1, the high dimensional data inputted into the SOM is denoted as the node in the input layer, and the two-dimensional output data of SOM is the neuron in the Kohonen layer with a weight vector $w$. The weight vector is used to denote the connection strength between the input node and the neuron.
For an input vector $x$, SOM will calculate the distance of $w$ of each neuron in Kohonen layer to the $x$, then identify the neuron with the shortest distance to $x$ as the winning neuron. Next, updating the $w$ of the winning neuron and its neighbour neurons with the equation (1).

$$\Delta W_i = C_i(X_t - W_{old}^i), \text{ for } i \in N_m$$

(1)

where $C_i$ is the gain function, $t$ is the index of the vector $x$ in the input data set, $N_m$ is the set of the winning neuron and its neighbors. The pseudo code of SOM is shown in table 2.

![Figure 1. The structure of the SOM.](image)

**Table 2. The pseudo code of SOM algorithm.**

```plaintext```
while (w is not stabilized) {
    obtaining an input vector $x$ form input layer;
    choosing random values to initialize the $w$;
    for (each of the neuron in Kohonen layer) {
        calculating the distance between its $w$ and the $x$;
    }
    the neuron with the shortest distance to the $x$ is selected as the winning neuron;
    updating the $w$ of the winning neuron and its neighbor neurons to reduce the distance between the $w$ and the $x$.
```

### 4. Experiment

The information of restaurants in Xi’an is collected by the web crawler and then analyzed as the input data by SOM. 8 attributes listed in table 1 make up the input vector of SOM. To reduce the complexity, some attributes are limited to a certain range of values. For instance, there are more than 70 addresses in the collected restaurant information, only the “address” attribute with frequency greater than 4% in all of the data records will be retained. Similarly, only the data records with the frequency of “type” attribute greater than 7% is used to train SOM. The clustering result is shown in figure 2, which indicates the segmentation of restaurants. Each figure in figure 2 indicates the distribution of a single attribute, while the blue color represents for the low value and red for high value.

All of the restaurants are clustered into 10 segments. Figure 2(a) to figure 2(d) illustrate the distribution of the overall rating, taste rating, environment rating and the service rating. It could be found that the distribution of high rating restaurants of these four attributes is roughly similar, mainly distributed in S1, S2, S3, S4, S5, S9 and S10. Specifically, the restaurants with the highest taste score are mainly distributed in S3, S1 and a bit of S4. There are also some restaurants with high environment rating distributed in S7 and S8. The distribution of low rating restaurants of these four attributes mainly distributed in S8 and S6. Figure 2(e) indicates that the restaurants with the highest per capita consumption distributed in S5 and S4, while the lowest mainly located in S1 and S6. Figure 2(f) shows the number of reviews in different segments. In general, the more reviews restaurants get, the more popular they are. The restaurants with the largest number of reviews distributed in S2. Figure 2(g) to Figure 2(l) illustrate the distribution of restaurant locations. For instance, restaurants located in the Bell tower/Drum tower mainly distributed in S2, and a little part of S1, S4, S5 and S7. Similarly, restaurants located in Qvjiang district, Xiaozhai and Gaoxin road are distributed in different segments according to
Different from these addresses, restaurants located in Weiyang road mainly distributed in only one segment: S10, so as the restaurants located in Tangyan road only distributed in S9. Figure 2(m) to figure 2(r) show the distribution of cuisine categories. The restaurants of hotpot mainly distributed in S2, S3 and S8; western restaurants mainly distributed in S4; restaurants of Shaanxi cuisine mainly distributed in S7; dessert shop mainly distributed in S1; restaurants of Japanese cuisine mainly in S5, and restaurants of snack food mainly in S6.

Through the clustering result of SOM, some representative features are summarized below.

S1: This is a segment of dessert shops. The overall and taste ratings of these shops vary widely, both good or bad, while the environment and service ratings are relatively good. The cost in these shops is low.

S2 and S3: Both of them are segments of hotpot restaurants. S2 mainly located in Bell tower\Drum tower and Qvjiang district, with the largest number of reviews in all of the segments, which indicates that the hotpot restaurants in these places are more popular than other restaurants. S3 mainly located in Xiaozhai and Gaoxin road. All of the four ratings of these two segments are relatively high.

S5: Japanese restaurants receive relatively good service and environment ratings, while the per capita consumption of this segment is the highest.

S6: Restaurants of snack food receive relatively poor reviews about overall ratings and taste ratings, while the environment and service ratings are good or bad. The consumption in these restaurants is one of the lowest in all the segments. It is worth noting that some restaurants in this segment are very popular but have low overall rating and taste rating, which mainly located in Bell tower\Drum tower.

S8: This is another segment about hotpot restaurants located in Qvjiang. These restaurants receive low overall rating, low taste rating. A few of these restaurants also perform poorly in terms of environment and service. However, the cost of this segment is not low.
5. Conclusion
In this study, an unsupervised learning approach named SOM is applied to the clustering task of restaurants. The online reviews of Xi’an restaurants in Dazhong Dianping are collected as the data source for model training. The clustering result of SOM indicates the following features of restaurants.

1. The dessert shops generally have good environment and low cost, but the tastes are uneven.
2. Some hotpot restaurants located in Bell tower/Drum tower are more popular than other restaurants, while some hotpot restaurants located in Qyjiang have poor word-of-mouth on overall rating, taste and service, however, the costs are not low.
3. Japanese restaurants usually offer better environment and service, but more expensive. Interestingly, some snack restaurants are very popular, but the overall rating and the test rating is low.
4. Snack restaurants are cheap, but the taste, service and environment are not guaranteed.
5. The feature of western restaurants is similar to Japanese restaurants. The expensive western restaurants are mainly located in Qyjiang district.

Acknowledgments
This work is supported by the Science Research Project of Shaanxi Provincial Department of Education (Grant No: 17JK0614) and the Youth Innovation Fund of Xian Shiyou University (Grant No: 2013BS025).
References

[1] Li, J., Xu, L., Tang, L., et al. (2018) Big data in tourism research: A literature review. Tourism Management, 68: 301-323.

[2] Kwok, L., Yu, B. (2013) Spreading social media messages on Facebook: an analysis of restaurant business-to-consumer communications. Cornell Hospitality Quarterly, 54(1): 84-94.

[3] Vu, H. Q., Li, G., Law, R., et al. (2018) Travel Diaries Analysis by Sequential Rule Mining. Journal of Travel Research, 57(3): 399-413.

[4] Xiang, Z., Du, Q., Ma, Y., et al. (2017) A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism. Tourism Management, 58: 51-65.

[5] Phillips, P., Zigan, K., Silva, M. M. S., et al. (2015) The interactive effects of online reviews on the determinants of Swiss hotel performance: A neural network analysis. Tourism Management, 50: 130-141.

[6] Liu, S., Law, R., Rong, J., et al. (2013) Analyzing changes in hotel customers’ expectations by trip mode. International Journal of Hospitality Management, 34: 359-371.

[7] Vu, H. Q., Li, G., Law R., et al. (2015) Exploring the travel behaviors of inbound tourists to Hong Kong using geotagged photos. Tourism Management, 46: 222-232.

[8] Li, G., Law, R., Vu, H. Q., et al. (2015) Identifying emerging hotel preferences using emerging pattern mining technique. Tourism management, 46: 311-321.

[9] Zhang, Z., Ye, Q., Zhang, Z., et al. (2011) Sentiment classification of Internet restaurant reviews written in Cantonese. Expert Systems with Applications, 38(6):7674-7682.

[10] Chou, C. L., Tsai, M. H., Chao, C. H., et al. (2014) Automatic restaurant information and keyword extraction by mining blog data for Chinese restaurant search. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining. Tainan, Taiwan. pp. 700-711.

[11] Park, S. B., Jang, J., Ok, C. M. (2016) Analyzing Twitter to explore perceptions of Asian restaurants. Journal of Hospitality and Tourism Technology, 7(4):405-422.

[12] Hwang, J., Zhao, J. (2010) Factors Influencing Customer Satisfaction or Dissatisfaction in the Restaurant Business Using AnswerTree Methodology. Journal of Quality Assurance in Hospitality & Tourism, 11(2):93-110.

[13] Claypo, N., Jaiyen, S. (2015) Opinion mining for Thai restaurant reviews using K-Means clustering and MRF feature selection. In: International Conference on Knowledge & Smart Technology. Chon Buri, Thailand. pp.105-108.

[14] Liao, H., Li, Y., Hu, T., et al. (2016) Inferring restaurant styles by mining crowd sourced photos from user-review websites. In: International Conference on Big Data. Alicante, Spain. pp. 937-944.

[15] Xie, D. X., Xia, W. F. (2013) Design and Implementation of the Topic-Focused Crawler Based on Scrapy. Advanced Materials Research, 850-851:487-490.

[16] Kohonen, T. (2013) Essentials of the self-organizing map. Neural Networks, 37(1):52-65.