D-020

Discovering Food Culture Between Different Countries Using Twitter Data

Mariem Akrimi† Yoshinori Miyazaki† Shohei Yokoyama‡

1. Introduction

In the past few years, there has been a huge growth in the use of microblogging platforms such as Twitter [1, 2], where a huge number of tweets with different topics and several locations are daily posted and where a good part of them is related to food. Nowadays, food has become one of the most popular topics discussed and shared on social media, therefore, the numbers of food data are hugely increased. Since food is a broad part of every culture around the world, eating choices can greatly reflect someone's passions beliefs and cultural identity. In this paper, we propose a method to discover food culture focusing on people food habits using twitter data. Twitter users often publicly express personal information messages like "I'm eating ramen", "I ate a whole pizza for lunch". Such a large number of similar tweets can be revealing, like discovering food festivals, meal time etc. Twitter streaming API is used to continuously collect tweets with a specific term like "Asagohan". We analyze the timestamp of the tweet to understand the eating time of users. Term frequency is also calculated, to extract word co-occurrence with each meal term.

2. Data Retrieval and System Architecture

2.1 Data Retrieval

By using Twitter streaming API [3], we collect a real-time tweet filtered by a specific term and a set of spatial bounding boxes defined by latitude and longitude of the tweet location. Data streamed by this API are JSON encoded, which is easy to track and to read. The first step of this study is to focus on twitter data only from Japan, and we assume that all the tweets are in the Japanese language. The set of tweets that we collect are based on a set of terms that match with the food topic.

2.2 System Architecture

Fig.1 describes a brief overview of the proposed system architecture. We crawl tweets that contain at least one of the following terms: "朝食", "ランチ", "夕食"," 美味しい","いただきます". Crawled tweets also include other metadata like the tweet ID, the timestamp, location etc. In this study, the parameters of interest are the tweet message, the user location and the time when the tweets were posted commonly known as a timestamp. Tweets are stored into MongoDB which is known to be a Relational Database Management System that provides a large variety of indexes including geo-indices which is very suitable for the query performance.

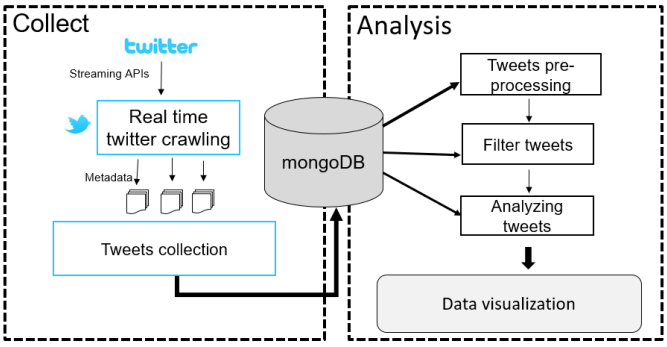

Fig.1: System architecture diagram.

3. Data processing and analysis

3.1 Data pre-processing

Pre-processing is a very fundamental data mining technique to transform raw data into an understandable format. Twitter data is highly susceptible to noise, missing values, and inconsistency that can affect the mining result, therefore in order to improve the quality of data as a first step, we clean the collected tweets from noisy tokens, characters, URL and stop words then, to simplify our work we divide the tweets into three different groups, where the first group contains only tweets with the term "朝食" (i.e.: breakfast in Japanese). The second group has only tweets with the term "ランチ" (i.e.: lunch in Japanese), and finally, the third group includes only tweets with the term "夕食" (i.e.: dinner in Japanese). We finish this step by tokenizing all tweet text into a set of meaningful pieces.

3.2 Analyze tweets based on time of the day

In this section, and after classifying the tweets by term, and in order to discover the specific time range of eating, we analyze the time of the posted tweets (i.e.: tweets timestamp). As twitter store all "Dates" and "Times" in the Coordinated Universal Time zone format mostly known as UTC time zone. The timezone and the geolocation (geo-tag) of the tweets are used to extract the exact timestamp of the tweet. We add 9 hours to the UTC timing for tweets geo-located in Japan, (i.e.: UTC+9). The histogram in Fig.2 shows the number of tweets about "朝食" as a function of the hours of the day. As stated in the Fig. 2, timing from 7:30 am to 8:30 am is the most popular

† Shizuoka University
‡ Tokyo Metropolitan University

time to tweet about “朝食” in Japan. This result shows that the most common time for having breakfast in Japan is within this timing interval.

As shown previously we also analyzed tweets about the term “夕食” as well. The histogram in Fig.3 shows the number of tweets related to the term “夕食” as a function of the hours of the days. This result states that the highest number of tweets is the time interval from 18:30 pm to 20:30 pm which is a perfectly reasonable result compared to the statistics given by the Ministry of Internal Affairs and Communications in 2011, according to their survey the average time of dinner is 19:06 pm on weekdays [4]

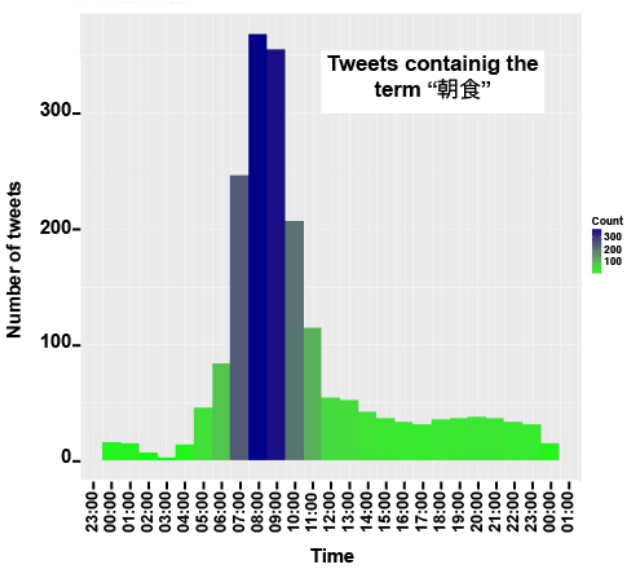

Fig.2: Tweets containing the term “朝食”.

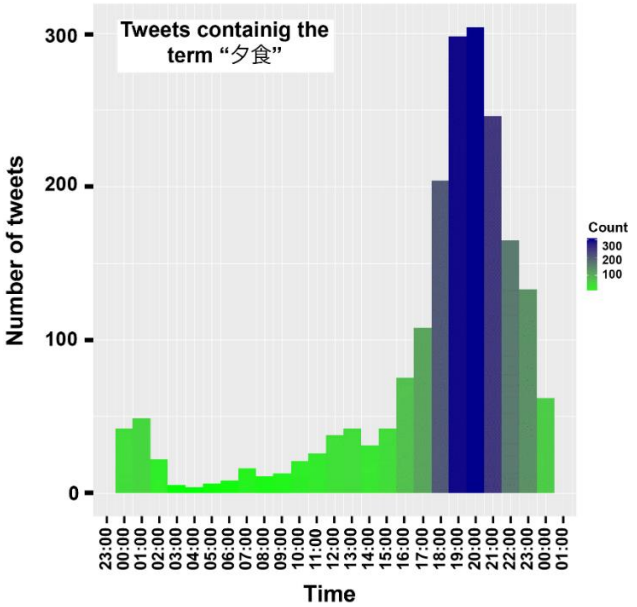

Fig.3: Tweets containing the term “夕食”.

3.3 Analyze tweets based on term frequency

In order to extract food-related term frequency, we first tokenize tweets using The Japanese morphological analyzer “Kuromoji” [5]. Data that includes the term “朝食” are transformed into a list of words, and then, we manually created a dictionary that contains a list of the most common Japanese food. Then, we extracted all related term matching with our list to calculate “朝食” tweets term frequency, from the result depicted in Fig.4, rice, miso soup, natto, pancake, and pizza are the most popular food consumed in the morning in Japan.

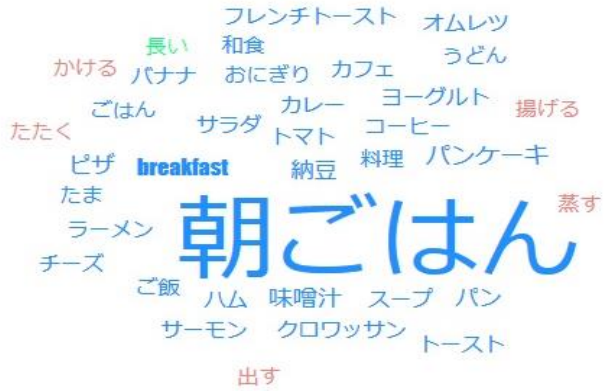

Fig.4: “朝食” Wordcloud.

4. Conclusion

In this paper, we have presented a research study based on Twitter data. We Analyzed people food habit focusing on the time of eating and the most common food eaten in each meal. The data analysis and results show that the most common time for breakfast and dinner in Japan are between 7:30 am to 8:30 am and between 18:30 pm to 20:30 pm respectively. The analysis shows also that rice, miso soup, pancake and natto are the most candidates for Japanese people breakfast. These results could be very interesting and challenging when compared to results from other countries.

References

[1] Stojanovski, D, Chorbev, I, Dimitrovski, I and Madjarov, G. 2016. Social Networks VGI: Twitter Sentiment Analysis of Social Hotspots. In: Capineri, C, Haklay, M, Huang, H, Antoniou, V, Kettunen, J, Ostermann, F and Purves, R. (eds.) European Handbook of Crowdsourced Geographic Information, Pp. 223–235. London: Ubiquity Press. DOI: http://dx.doi.org/10.5334/bax.q. License: CC-BY 4.0.

[2] H. Achrekar, A. Gandhe, R. Lazarus, Ssu-Hsin Yu and B. Liu, "Predicting Flu Trends using Twitter data," 2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), Shanghai, 2011, pp. 702-707.

[3] https://developer.twitter.com/content/developer-twitter/en.html, accessed on March 18, 2018.

[4] https://www.e-stat.go.jp
http://www.atilika.org/