How to Use the Social Media Data in Assisting Restaurant Recommendation

Wenjuan Cui[1], Pengfei Wang[1,2], Xin Chen[1], Yi Du[1], Danhuai Guo[1],
Yuanchun Zhou[1](✉), and Jianhui Li[1]

[1] Computer Network Information Center, Chinese Academy of Sciences,
Beijing 100190, China
zyc@cnic.cn

[2] University of Chinese Academy of Sciences, Beijing, China

Abstract. Online social network applications such as Twitter, Weibo, have played an important role in people's life. There exists tremendous information in the tweets. However, how to mine the tweets and get valuable information is a difficult problem. In this paper, we design the whole process for extracting data from Weibo and develop an algorithm for the foodborne disease events detection. The detected foodborne disease information are then utilized to assist the restaurant recommendation. The experiment results show the effectiveness and efficiency of our method.

Keywords: Recommender system · Event detection · Social media · Foodborne disease

1 Introduction

With the development of information technology, people spend more time surfing on the internet. While people's lives get convenient from the information services, the continuously generated huge volume of data make it difficult to easily get the useful information fulfilling people's requirements. Recommender system was designed aiming to overcome the information overload problem [1]. Except for the traditional recommender systems, the personalized recommendation has been developed and applied in various fields to meet the users' interest [2]. Lots of approaches have been proposed to the recommendation research, in which content based approach, collaborative filtering and hybrid models are most used methods [3–5]. Various recommender systems have been deployed in tremendous applications. However, most of the systems only consider one particular data source. They seldom care about the information from other data sources like the social network.

Twitter, Facebook and other social network applications have been frequently used in recent years. People tend to express their opinions, feeling or just tell friends what they are doing on the social network. Twitter is a popular microblogging service which attracts much attention. People may post tweets at any place and during any time. In general, the length of one tweet has a limit of 140 characters. There may be only little information in a particular tweet, but the

H. Gao et al. (Eds.): DASFAA 2016 Workshops, LNCS 9645, pp. 134–141, 2016.
DOI: 10.1007/978-3-319-32055-7_12

accumulated content can generate a vast amount of information and include important knowledge. The context of the tweet also provides lots of information [6]. Twitter has helped to provide valuable message for various applications. Tumasjan et al. tracked the public political opinions on Twitter and predict the election result [7]. The stock could also be predicted [8]. Sakaki et al. investigated the real-time interaction of earthquakes in Twitter and proposed a method to monitor the tweets and to detect the earthquakes. They can detect the earthquakes with high probability and faster than the government [7].

Tweets can also be used in the area of public health. Users often post the messages like “I got a flu, getting a running nose” or “got a stomachache after eating pizza”. Millions of such messages may give a direction for the influenza tracking or other public health problems. Aramaki et al. propose a machine learning method to detect the influenza epidemics from twitter data [9]. Tweets can be used to track the influenza and forecast future influenza rates with high accuracy [10,11]. Twitter data have been used on surveillance of other public health related problems like Dengue and foodborne disease [12].

In this paper, we will consider the relevance of foodborne disease and the restaurants with contaminated food. First, we will crawl the foodborne disease related tweets from Weibo, which is a Chinese social network application similar to Twitter. Then the tweets are classified to filter the part which are the actual foodborne disease related ones. The key-phrases are extracted and a SVM classifier is designed to detect the foodborne disease events. Meanwhile, the locations are also determined and the restaurants with contaminated food are identified. Finally the foodborne disease factor is used in the restaurant recommendation.

2 Related Work

Foodborne disease (or foodborne illness) refers to any disease resulting from the consumption of contaminated food. Foodborne Disease Outbreak (FBDO) is defined as the two or more cases of a similar illness resulting from the ingestion of a common food. According to CDC 2011 Estimates, each year roughly 1 in 6 Americans (or 48 million people) gets sick, 128,000 are hospitalized, and 3,000 die of foodborne diseases [13]. In the past years, the tracking and detection of foodborne disease were mainly carried on by surveillance systems. But the traditional surveillance systems have the limit of time lag in the detection of FBDO. Recently, the social media data have been introduced to the surveillance of foodborne disease [14–16].

The users for Twitter or Weibo get more and more and the functions of these social network applications get complex. People tend to express their feelings and describe what they are doing on the social network applications. The event detection from the short text data such as tweets has been of significant value. The influenza and foodborne disease could also be detected by the Twitter or Weibo data. The response time is proved to be faster than the traditional surveillance systems.

As the foodborne diseases are mostly caused by contaminate food, the twitter data can also help with the identification of restaurants related with foodborne diseases. Sadilek et al. combine the Twitter data with foodborne disease and restaurants [17]. They collect the tweets posted in mobile devices and get the geo-location of the tweets, and then find the tweets which are near to some restaurants. After that, they design a human guided machine learning method to classify the foodborne disease events. Finally, the detected tweets are associated with the restaurants. Their application is deployed in Las Vegas [18]. Their method performs well comparing with the health department. However, they only collect the mobile data with specified geo-locations. For the data which do not contain specific geo-locations, their method does not work. In this paper, we will design a method to solve this problem.

The paper is organized as follows. Section 1 introduces the background and Sect. 2 shows some related works. In Sect. 3, the process for the foodborne disease event detection will be illustrated step by step. Section 4 will show the experiment results and we will conclude in Sect. 5.

3 The Method

3.1 Data Filtering

To get the relevance of foodborne diseases and the restaurants with contaminated food, we should first get the tweets with the information of foodborne diseases. There is a government surveillance system for the foodborne diseases which collects foodborne disease cases in 13 provinces in China from 2013. Each case is reported by the doctors from the sentinel hospital. We analyze the symptoms of these cases and get a list of keywords for them. The frequency of appearance for each keyword is calculated and the keywords with high frequency are selected.

With the help of Weibo API, we then crawl the tweets in Weibo which contain the keywords of the symptoms for the foodborne diseases. While the tweets contain huge amount of information, there are also lots of noisy data. The extracted tweets may be from the public accounts which offer health advices. To reduce the effect of the useless tweets, we build a support vector machine (SVM) classifier to filter the tweets, which is based on some properties of the tweets and the users. After the Weibo data are filtered, we should identify the foodborne disease event.

3.2 Foodborne Disease Event Detection

As the contents of tweets are texts, we should first convert the natural language into mathematical format in order to process the tweets using machine learning methods. Constructing the vector representation for the words is a good way to capture the syntactic and semantic word relationships. The toolkit word2vec is an open source software developed by Google, which is used to convert the words into vectors. It is efficient while billions of words could be trained in a day, which makes it possible for us to train our data from Weibo. We construct the vector representations for the words in tweets using word2vec. Then the sematic similarities between the tweets could be calculated by the similarities in the vector space [19,20].

There are huge amount of data continuously generated in Weibo and the topics change fast. Except for the tweet containing the keywords for foodborne diseases, the tweets in its context may contain other important information for the foodborne diseases, such as the location where the foodborne diseases are caused. Assuming the tweets for a user is a tweet list $S = \{T_1, T_2, ..., T_k, ..., T_n\}$, where T_k is the tweet which contains the keywords for the foodborne disease symptoms and T_i is posted earlier than T_j if $i < j$.

We design an algorithm to dynamically choose more relevant tweets in the context to get a larger candidate corpus. We carry on a tokenization for the tweets and construct the vector representation for each tweet. For two tweets T_i and T_j, the vectors for them are v_i and v_j respectively. We define the semantic similarity between two tweets as the cosine similarity of their word vectors. The equation is illustrated in formulas (1).

$$Sim(T_i, T_j) = \cos(v_i, v_j) \tag{1}$$

The similarity between two tweets gets higher while the value of cosine is larger. We compute a dynamic window in the context for a particular tweet T_k which contains the keywords for the foodborne disease symptoms. The algorithm is shown as Algorithm 1.

For a tweet T_i, we compute the similarity between T_k and T_i. If the similarity is greater than the threshold U, the tweet T_i will be added into the candidate tweet set C. We find the similar tweets with T_k before and after it. Our method takes account of the semantic similarity between the tweets and makes sure that the tweets in the candidate set are relevant to the foodborne disease. At the same time, less noisy data are introduced.

After the candidate tweet set is built, we try to extract the key-phrases for the tweets. The easiest method is based on TF/IDF. But it only considers the statistical properties of the phrases. The relationships between the phrases are not considered and the phrases with low frequency will be ignored. In this paper, we use TextRank, which is a graph based key-phrase extraction algorithm [21]. It divides the text into several segments and builds the graph model for them. The voting schema is used to rank the phrases of the text and the key-phrases are extracted according to the rank.

We manually label some tweets which are indeed related with foodborne disease. And then we use the extracted key-phrases for the tweets together with the keywords for the symptoms of the foodborne disease to train a SVM classifier to detect the foodborne disease event.

3.3 Restaurant Recommendation

When the foodborne disease events are detected, we want to associate the results with the restaurant recommendations. To find the restaurants which are related

Algorithm 1. Dynamic context window calculation

```
Input: A tweet list for a user $S = \{T_1, T_2, ..., T_k, ..., T_n\}$; The tweet $T_k$ which contains
  the keywords for the foodborne disease symptoms; The threshold $U$ for the similarity
  between tweets; The decreasing rate for the similarity measure $\eta$; The upper bound
  $P$ for the dynamic window and the lower bound $Q$ for the dynamic window.
Output: The candidate tweet set C
  Initialize $T = T_k$,$C = \emptyset$
  Push $T_k$ into C
  for i=1 to P do
    if $Sim(T, T_{k-i}) > U$ then
      Push $T_{k-i}$ into C
      Update $T = T + T_{k-i}$,Update $U = U * \eta$
    else
      break;
    end if
  end for
  Update $T = T_k$
  for j=1 to Q do
    if $Sim(T, T_{k+j}) > U$ then
      Push $T_{k+j}$ into C
      Update $T = T + T_{k+j}$,Update $U = U * \eta$
    else
      break;
    end if
  end for
  return C
```

with the foodborne diseases, the geo-location for the foodborne disease event should be determined. There may be location description in the registration information for the Weibo users. But this location is the zone where the user lives, not exactly the location of the foodborne disease event occurs. There are also GPS data for some mobile users, but the data are also sparse. On the other hand, we observe that lots of the tweets related with foodborne disease contain the information of the restaurant or the name of the food. And some users also refer to the location when they post a tweet. We utilize this information with the aid of other data to find the restaurants related to the foodborne disease.

We define the restaurant and food information in the tweet as information A. If the tweet contains the restaurant, we will get the location of the restaurant from the website of dianping(https://www.dianping.com/), which has the information for the restaurants and their detailed locations. If the tweet contains the name of the food, we could find all the restaurants with this food and their location from Baidu API. The tweets may also contain the geo-location when they are posted. We define this kind of information as information B. We will use the administrative location data on tcmap(http://www.tcmap.com.cn) to get the detailed location for information B. The location in the registration information of the Weibo users are defined as information C. Note that information A and

B may be missed, but information C is usually available. We then design an algorithm which utilizes the information A, B and C to get the location L of the foodborne disease events. The location L is determined as in formulas (2).

$$\begin{cases} L = \{A_i | minDist(A_i, B_j), A_i \in A, B_j \in B\}, & A \neq \emptyset, B \neq \emptyset \\ L = \{A_i | A_i \in C, A_i \in A\}, & A \neq \emptyset, C \neq \emptyset, B = \emptyset \\ L = \{B_j | B_j \in C, B_j \in B\}, & B \neq \emptyset, C \neq \emptyset, A = \emptyset \end{cases} \quad (2)$$

When A and B are both available, we compute the nearest two points $A_i \in A$ and $B_j \in B$ and A_i is the desired location. When A and C are available and B is empty, the locations in A which are also in C are the desired locations. When B and C are available and A is empty, the locations in B which are also in C are the desired locations.

After the locations for the foodborne disease events are determined, we find all the restaurants which are related with the foodborne disease events. We give a score for each restaurant and the score is lower for a restaurant related with the foodborne diseases. Then we insert the foodborne disease related information as a factor in the restaurant recommendation algorithm.

4 Experiments

In the experiments, we extract the tweets from 933,313 users in Beijing, China which contain the 31 keywords for the symptoms of foodborne diseases. The tweets are posted between August 2014 and October 2014. We construct the word vectors for the tweets using word2vec, and then use Algorithm 1 to get the candidate tweet set. There are totally about 80 million tweets in the set.

We use the ten features(#followings, #followers, length of personal description, #all tweets, #average retweeting, #average recommendation, #average comments, average length of the tweets, time of posting, # average links in tweets) to construct a SVM classifier to filter the tweets. After data filtering, we get about 31 % of the tweets in the candidate tweet set.

Except for Algorithm 1, there is also a method which selects a fixed number of tweets before and after the tweet T_k which is foodborne disease related. We select the 200 tweets before and after T_k into the candidate tweet set and compare it with our proposed Algortithm 1.

Table 1. Comparison of fixed context window calculation and Algorithm 1

Method	Accurate rate
Fixed context window calculation	27.4 %
Dynamic context window calculation	39.2 %

From Table 1 we can see that our proposed algorithm can get the candidate tweet set semantically and get higher accurate rate for the foodborne disease event detection.

We make statistics to the information A, B, C mentioned in Sect. 3.3. There are 13 % of all the tweets which contain both A and B, 19 % containing A and C but not B, and 16 % containing B and C but not A. For the case where A, B and C are all available, we calculate the accurate rate for the location determination. The result is shown in Table 2.

Table 2. The accurate rate for location determination

Total number	Correct number	Accurate rate
500	332	66.4 %
1000	647	64.7 %
1500	1009	67.3 %
2000	1280	64.0 %

From all the above experiment results, we see that the accurate rate for location determination and event detection are not high. That is partly caused by the characteristics of the tweets.

5 Conclusion

In this paper, we show how to use the social media data to extract valuable information about foodborne diseases. Algorithms are designed to get the foodborne disease related tweets and detect the foodborne disease events. The experiments show that our methods are effective. However, the accurate rate for the event detection is not high due to the sparsity and continuously changed topics of Weibo. In the future work, we will try to improve the detection algorithm and use the algorithm to assist the restaurant recommendation.

Acknowledgments. This work was supported by the National Natural Science Foundation of China under Grant No. 61402435,41371386,91224006, and the Knowledge Innovation Program of Chinese Academy of Sciences under Grant No. CNIC_QN_1507.

References

1. Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Trans. Knowl. Data Eng. **17**(6), 734–749 (2005)
2. Sharma, L., Gera, A.: A survey of recommendation system: research challenges. Int. J. Eng. Trends Technol. (IJETT) **4**(5), 1989–1992 (2013)
3. Koren, Y., Bell, R., Volinsky, C.: Matrix factorization techniques for recommender systems. Computer **8**, 30–37 (2009)
4. Su, X., Khoshgoftaar, T.M.: A survey of collaborative filtering techniques. Adv. Artif. Intell. **2009**, 4 (2009)

5. Shi, Y., Larson, M., Hanjalic, A.: Collaborative filtering beyond the user-item matrix: A survey of the state of the art and future challenges. ACM Comput. Surv. (CSUR) **47**(1), 3 (2014)
6. Xie, H., Li, Q., Mao, X.: Context-aware personalized search based on user and resource profiles in folksonomies. In: Sheng, Q.Z., Wang, G., Jensen, C.S., Xu, G. (eds.) APWeb 2012. LNCS, vol. 7235, pp. 97–108. Springer, Heidelberg (2012)
7. Tumasjan, A., Sprenger, T.O., Sandner, P.G., Welpe, I.M.: Predicting elections with twitter: What 140 characters reveal about political sentiment. In: ICWSM 2010, pp. 178–185 (2010)
8. Li, X., Xie, H., Song, Y., Li, Q., Zhu, S., Wang, F.: Does summarization help stock prediction? news impact analysis via summarization. IEEE Intell. Syst. **30**(3), 26–34 (2015)
9. Aramaki, E., Maskawa, S., Morita, M.: Twitter catches the flu: detecting influenza epidemics using twitter. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1568–1576. Association for Computational Linguistics (2011)
10. Signorini, A., Segre, A.M., Polgreen, P.M.: The use of twitter to track levels of disease activity and public concern in the US during the influenza A H1N1 pandemic. PloS ONE **6**(5), e19467 (2011)
11. Culotta, A.: Detecting influenza outbreaks by analyzing twitter messages (2010). arXiv preprint arXiv:1007.4748
12. Gomide, J., Veloso, A., Meira Jr., W., Almeida, V., Benevenuto, F., Ferraz, F., Teix-eira, M.: Dengue surveillance based on a computational model of spatio-temporallocality of twitter. In: Proceedings of the 3rd International Web Science Conference, p. 3. ACM (2011)
13. Center for Disease Control Prevention (CDC): CDC estimates of foodborne illness in the United States. Retrieved 23 March 2011
14. Newkirk, R.W., Bender, J.B., Hedberg, C.W.: The potential capability of social media as a component of food safety and food terrorism surveillance systems. Foodborne Pathog. Dis. **9**(2), 120–124 (2012)
15. Harris, J.K., Mansour, R., Choucair, B., Olson, J., Nissen, C., Bhatt, J., Brown, S.: Health department use of social media to identify foodborne illness-chicago, illinois, 2013–2014. MMWR Morb. Mortal Wkly. Rep. **63**(32), 681–685 (2014)
16. Xie, H., Yu, L., Li, Q.: A hybrid semantic item model for recipe search by example. In: 2010 IEEE International Symposium on Multimedia (ISM), pp. 254–259. IEEE (2010)
17. Sadilek, A., Brennan, S., Kautz, H., Silenzio, V.: nEmesis: Which restaurants should you avoid today? In: First AAAI Conference on Human Computation and Crowd- Sourcing (2013)
18. Sadilek, A., Kautz, H., DiPrete, L., Labus, B., Portman, E., Teitel, J., Silenzio, V.: Deploying nemesis: Preventing foodborne illness by data mining social media (2016)
19. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space (2013). arXiv preprint arXiv:1301.3781
20. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013)
21. Mihalcea, R., Tarau, P.: Textrank: Bringing order into texts. Association for Computational Linguistics (2004)