Int. J. Nonlinear Anal. Appl. 15 (2024) 10, 157–162
ISSN: 2008-6822 (electronic)
http://dx.doi.org/10.22075/ijnaa.2023.31572.4670

Predicting the approximate time of event occurrence based on changes in the speed of sending messages in X social network

Abulfazl Yavari

Faculty of Computer Engineering and IT, Payame Noor University, Tehran, Iran

(Communicated by Seyyed Mohammad Reza Hashemi)

Abstract

In recent years, the availability of virtual social network data and the mutual impact that real and virtual communities have on each other has led to many research in the field of virtual social network analysis. Detecting and predicting the occurrence of social events is one of the important applications of this field. In this paper, using a threshold structure, the approximate time of the event is predicted by analyzing the messages of social network X (former Twitter). In the proposed method, the data is first partitioned and preprocessed and then clustered using the distance-based Chinese restaurant process. Changes in the speed of sending messages to each cluster are used as an effective feature in predicting the approximate time of an event. Experiments conducted on almost 5 million tweets including 876 events show a prediction accuracy of 78%.

Keywords: Virtual social network analysis, Social network X (former Twitter), Distance-based Chinese restaurant process, Message sending speed changes
2020 MSC: 91D30

1 Introduction

A social network is actually a social structure that consists of nodes that are generally individual or organizational, and these nodes are connected based on one or more specific dependencies, such as ideas, financial exchanges, friends and relatives, web connections, or disease transmission [12]. Virtual social media are applications based web 2. This means that users are able to create content themselves, organize and adjust it, share their information and content with others, or criticize and change it. Data generated in social networks are valuable because they reflect different aspects of real-world communities. On the other hand, these data are easily available through web crawlers or public APIs. These two features are the most important reasons why researchers study and investigate virtual social networks [9].

The event in the social network is actually a factor that creates significant changes in some parameters and characteristics of the social network. In recent years, most of the researches that have been done in the field of events by analyzing social networks have dealt with detecting and identifying the event [1, 8, 4]. Most of these methods try to predict the desired event after filtering tweets based on a sets of keywords and then performing statistical analysis. In this research, the approximate time of the event is predicted based on the changes in the speed of sending related messages.

Email address: a.yavari@pnu.ac.ir (Abulfazl Yavari)

Received: August 2023 *Accepted:* October 2023

As shown in [13], before the event, signs of the event can be seen in social networks, and the closer the event is, the number of related messages increases. In this article, based on these changes in the speed of sending messages to the social network, the approximate time of its occurrence is predicted.

In the continuation of the paper, first in section 2, a review of related works in the field of event prediction has been done. In section 3, the proposed method for predicting the approximate time of the event is given, and in section 4, the results of the experiments on the data set are given.

2 Related works

In this section, the works related to the proposed system for predicting the approximate time of the event are presented. First, the works related to event prediction are mentioned, and then the article [13], which is actually a base for the current paper, is presented in detail. In [5], a system for predicting the spread of Covid-19 based on the detection and tracking of events on Twitter is proposed. This system first creates knowledge graphs based on Twitter streams related to Covid-19. In other words, based on the new information and news that is published on the network about the identification of new cases of infection in different locations, the desired graph is updated, and then based on that, the prediction of how the disease will spread is done.

The authors in [6] predicted the sales of movies using the Least Squares Support Vector Regression (LSSVR) model. They applied the mentioned model on three different types of data, including movie database data such as Box Office and IMDB, Twitter data about movies, and mixed data. The results show that applying the model to the combined data obtains more accurate results than other data.

In [3], by analyzing blogs, they found a positive correlation between the volume of user posts about a particular singer and album and the sales of that music album. The authors in [11] also obtained such a correlation between Twitter messages and music album sales by analyzing Twitter. They used the messages that contained the name of the singer or the name of the album in the period of two weeks before the start of sales until one week after it.

Burnap et al. [2] proposed a model based on sentiment analysis to predict the outcome of the 2015 UK election. Based on the method of sentiment analysis presented in [10], they assigned a score between -5 and +5 to each tweet, respectively, meaning a strong negative feeling to a strong positive feeling, and based on the sum of the feeling scores for each party; they predicted the number of party seats in the British Parliament.

In [14] and [15], two studies based on indicators are presented in order to predict the outcome of the American elections. In both studies, based on an exponential averaging, the amount of indicators is calculated for each party. Any party that gets a bigger indicator is expected to win the election.

In the paper [13] written by the author of this paper, in the first step, tweets are first categorized based on a time window with a fixed length, and then pre-processing is applied to clean the data on each category. In the second step, tweets are clustered. For this purpose, NMF algorithm has been used to create initial clusters and value them. After the initial clusters are created using only the first group of data, the next groups of data are sequentially clustered based on the ddCRP algorithm. During the clustering process, the amount of changes in the size of each cluster is monitored, and if they are within the desired threshold limits, the k most frequent words of the desired cluster are sent to the output as a description of the predicted event. The mentioned article does not provide any information about the approximate time of the event and only predicts the occurrence of an event in the future. But in the current article, using a multilayer artificial neural network, the approximate time of the event is determined.

3 Proposed Method

In this section, the proposed method for predicting the approximate time of an event is described. In order to make a forecast with proper accuracy, the effective parameters in the forecast should be accurately identified and monitored and evaluated. Among the features that can be evaluated in the event detection and prediction process are the registration rate of new users, the speed of sending messages to the network, the number of hashtags, the number of retweets, the number of mentions, the number of links, the number of user responses to messages, and the rate of sending text. They are effective by people.

In this paper, the speed of sending messages to the network is used. The goal is, by calculating the speed of sending messages, and comparing them with a series of thresholds that have been obtained experimentally, the approximate time of the occurrence of the event is type A (occurrence of more than one month), type B (occurrence of more than

one week to one month), type C (occurring between two days to one week) and type D (occurring less than two days) should be specified.

It was shown in [13] that before predictable events occur, messages about that event are sent on the network. In this research, with further investigations on the available data, it was found that the speed of sending messages of each event usually follows the following four-step pattern (Figure 1)

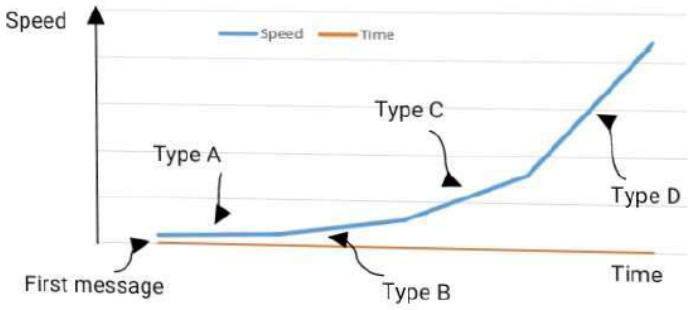

Figure 1: A four-step pattern to the occurrence of events

It should be noted that the events will not necessarily include the above four steps. An event may be noticed from one week to the occurrence in the social network. But what matters here is the rate of change in the speed of sending messages about that event, which shows the number of days until the event.

The initial step that appears with the first signs of the event, changes in the speed of sending messages (the slope of the above graph) is almost constant and lower than threshold L1. The slope of the graph is also calculated based on the Equation 1.

$$S = \frac{V_d - V_o}{d} \tag{1}$$

In Equation (1), V_d is the current speed, V_o is the speed of the first day, and is the number of days that have passed since the first sign occurred. Please note that you cannot use speed value instead of speed value changes. In fact, having a high speed does not indicate being close to the time of occurrence, because some events are inherently important and popular and are discussed a lot in the social network. For example, although there is a lot of time left for the World Cup event, the sending speed is high in the first step, but the changes in this speed do not change much in the following days. Therefore, by considering speed changes as a decision criterion, it will be possible to determine the remaining time until the occurrence of the event for less important and important events in the form of the same thresholds.

In the second step, changes in message sending speed are in the range of L1 and L2 and indicate near-term events. In the third step, changes in the speed of sending messages are in the range of L2 and L3 and show very close events. Finally, the final step shows the events whose growth changes are greater than L3 and are defined as immediate events.

The architecture of the proposed method is shown in Figure 2, consisting of three main parts. The first part, called pre-processing, includes two steps. First, incoming tweets are grouped based on a time window of fixed length. Then, the usual operation of text data pre-processing is performed in data mining systems in order to clean them, and finally the TF-IDF weight vector is created. In the second part, while clustering using two algorithms, NMF and ddCRP, the characteristics of changes in the speed of entering tweets into each cluster are calculated so that the approximate time of an event can be predicted.

Finally, in the third section, the predicted events are displayed along with the approximate time of occurrence. In the rest of this section, each of the main parts of the proposed method is explained in more detail.

A) Preprocessing

In general, in most text documents, there may be some useless data that, in addition to negatively affecting the final accuracy, also affect the volume of data and processing time. Therefore, it is necessary to apply the cleaning action on the documents before processing them. In this research, tweets are categorized based on a time window, the length of which can be adjusted at the start of the system, before data cleaning.

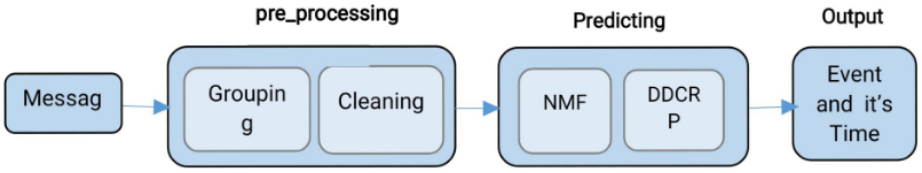

Figure 2: General architecture of the proposed method

After grouping the tweets based on their publication time, each group is read in order and the usual cleaning and pre-processing actions are performed on the data such as breaking the text string, removing stop words, removing numbers and links, and extracting bin words on the tweets. At the end of this stage, using the TF-IDF method (Equation 2), the matrix of document weight × word is created, which contains the importance of each word in the tweets, and is used as an input for the prediction stage.

$$TF - IDF(w, t, N, n_w) = f_{w,t} - log(\frac{N}{n_w}) \tag{2}$$

In Equation 2, $f_{w,t}$ is the frequency of word w in tweet t, N is the total number of tweets, and n_w is the number of tweets in which the desired word was present.

B) Predicting the time of occurrence of the event

In this research, incremental clustering method is used. At the beginning of the incremental clustering process, the first group of tweets are clustered based on the NMF algorithm. In this way, firstly, a good initialization is created for the initial clusters and their content, and secondly, a suitable initial structure is created faster in terms of time. Then, by reading each new set of tweets, the incremental clustering routine is executed. ddCRP incremental clustering method is used in this system. At the end of each step of incremental clustering, a set of clusters is available. In fact, each cluster represents an event. After clustering each group of tweets, the rate of entering tweets into each cluster is calculated and stored. Of course, since the grouping of tweets in the pre-processing stage was based on a time window with a fixed length, the rate of tweets entering in time interval $i, (V_{t,i})$, with the number of tweets entering each cluster in this interval ,$n_{t,i}$, will have a direct relationship (Equation 3).

$$V_{t,i} = \frac{n_{t,i}}{\parallel i \parallel} \tag{3}$$

In Equation 3, $\parallel i \parallel$ is the length of the time interval. Now, it is possible to predict the approximate time of occurrence of the event in the future based on the amount of changes in the rate of tweets entering each cluster and comparing this amount of changes with a series of threshold limits that are obtained experimentally.

C. Output of the proposed method

The last part of the proposed system is the visualization of events that are predicted as future events in the proposed method. From each cluster representing future events, the k most frequent words are output as event descriptions, as well as the approximate time of occurrence. (Table 1)

Table 1: EOutput of the proposed system as predicted events as their occurrence times

row	Six most frequent multi-cluster words as descriptions of predicted events	Type of occurrence time
1	Chemical, Kerry, use, house, weapons, stand	A
2	Divorce, old, request, wife, George, make	A
3	C Fire, hospital, immediate, dead, sandy, topic	C
4	Merkel, exit, angel, third, car, stop	D
5	Mayor, healthcare, new, india, first, york	B

4 Experiments and results

In this section, the tests performed on a dataset of tweets are given. First, the data set is introduced, and then the process of extracting the thresholds is described, and finally, the proposed method is evaluated based on the threshold limits of changes in tweet sending speed.

A. Dataset

The dataset used in this research was selected from the article [7]. This data set includes tweets taken from famous news media such as CNN, BreakNews, and BBC on the Twitter social network in a period of six months. The dataset contains approximately 43 million tweet IDs that refer to 5234 events. Each event in the dataset is described by a number of keywords that are used to evaluate the proposed system. But due to reasons such as the blocking of some user accounts and the deletion of events that had very few tweets, the final dataset included 1940 events and almost 12 million tweets. The statistical summary of the final data set is given in Table 2.

B. Determining the Thresholds

In order to determine the threshold limits based on which it is possible to predict the approximate time of the outbreak, 60% of the total events of the data set along with their related messages have been separated. In this way, 1074 events along with almost 7 million tweets have been evaluated (Table 2).

Table 2: Statistical summary of the test and train data set

	Dataset Statistics	Minimum	Maximum
Training	Number of tweets per event	1764	29767
	Number of keywords per event	2	29
Testing	Number of tweets per event	1120	38436
	Number of keywords per event	3	32

In the next step, the maximum changes in the speed of sending messages related to each event in four classes A (more than a month to occur), B (more than a week and less than a month), C (more than three days and less than a week) and d (less than three days) is calculated. Table 3 shows a sample of extracted information for 4 different events.

Table 3: Threshold values for 4 different type of events

	Class A	Class B	Class C	Class D
Event 1	0.1	0.69	1.32	5.36
Event 2	0	0.72	1.27	4.89
Event 3	0	0	1.84	5.72
Event 4	0	0	0	7.39

In Table **??**, sample events have been selected to cover the different modes of events in the dataset. Some events don't necessarily have a sign on the social network a month before they happen (events 2, 3 and 4) or some events appear on the network only a few days before they happen (Event 4).

The threshold limit of each class is calculated based on the pruned average. For each class, 5% of the edge data are removed to reduce the destructive effect of the highest and lowest numbers on the average, and then averaging is done. Therefore, the average maximum speed of sending messages in each class is determined as the threshold of that class.

C. The accuracy of the prediction

Now, based on the obtained threshold limits, the rest of the data is evaluated. The number of test data includes 876 events and almost 5 million tweets. According to the proposed method described in section 3, the data are first divided and then clustered based on the ddcrp clustering algorithm. Since each cluster represents an event, the changes in the speed of sending messages to each cluster in each time window are calculated and compared with the threshold limits. If they are within the desired threshold range, the K most frequent words of each cluster are sent to the output along with the predicted class. If the speed of sending messages to a cluster decreases for several time windows, and according to previous calculations, it is expected to happen, the event can be considered finished. Also, if a cluster remains for a period of time longer than the head in class A (more than 30 days of occurrence), it can be identified as noise and removed.

To calculate the accuracy of the proposed method in predicting the approximate time of the event, the Equation 4 is used.

$$Accuracy = Number\ of\ correct\ predictions / Number\ of\ events \tag{4}$$

The accuracy of the proposed method on the test data is calculated to be 78%.

5 Conclusion and future works

In this article, a method based on several thresholds was proposed to predict the approximate time of an event in the future. It was also shown that, usually, events experience different speed changes in terms of sending messages on the social network, depending on how much time is remained until they occur. Events often show a similar behavior in terms of changes in the speed of sending messages to the social network. In the continuation of this research path, the effect of categorizing events according to their type such as sports, social and political on the accuracy of prediction can be investigated. In addition, by extracting statistical features from each cluster and applying them to machine learning models, it is possible to provide a more adaptable and accurate method.

References

[1] H. Becker, M. Naaman, and L. Gravano, *Beyond trending topics: Real-world event identification on Twitter*, Proc. Int. AAAI Conf. Web Soc. Media **5** (2011), no. 1, 438–441.

[2] P. Burnap, R. Gibson, L. Sloan, R. Southern, and M. Williams, *140 characters to victory?: Using Twitter to predict the UK 2015 general election*, Electoral Stud. **41** (2016), 230–233.

[3] V. Dhar and E.A. Chang, *Does chatter matter? The impact of user-generated content on music sales*, J. Interact. Market. **23** (2009), no. 4, 300–307.

[4] W. Dou, X. Wang, W. Ribarsky, and M. Zhou, *Event detection in social media data*, IEEE VisWeek Workshop on Interactive Visual Text Analytics-Task Driven Analytics of Social Media Content, 2012, pp. 971–980.

[5] X. Fu, X. Jiang, Y. Qi, M. Xu, Y. Song, J. Zhang, and X. Wu, *An event-centric prediction system for COVID-19*, IEEE Int. Conf. Knowledge Graph (ICKG), 2020, pp. 195–202.

[6] Y.T. Huang and P.F. Pai, *Using the least squares support vector regression to forecast movie sales with data from Twitter and movie databases*, Symmetry **12** (2020), no. 4.

[7] J. Kalyanam, M. Quezada, B. Poblete, and G. Lanckriet, *Prediction and characterization of high-activity events in social media triggered by real-world news*, PloS one **11** (2016), no. 12, e0166694.

[8] A. J. McMinn, Y. Moshfeghi, and J.M. Jose, *Building a large-scale corpus for evaluating event detection on Twitter*, Proc. 22nd ACM Int. Conf. Inf. Knowledge Manag. 2013, pp. 409–418.

[9] N. Panagiotou, I. Katakis, and D. Gunopulos, *Detecting events in online social networks: Definitions, trends and challenges*, Solving Large Scale Learning Tasks, Challenges and Algorithms. (2016) 42–84.

[10] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, *Sentiment strength detection in short informal text*, J. Amer. Soc. Inf. Sci. Technol. **61** (2010), no. 12, 2544–2558.

[11] R. Vossen, *Does chatter matter? Predicting music sales with social media*, (2013), [online] Available at: https://www.basicthinking.de/blog/wpcontent/uploads/2013/06/Does-Chatter-Matter.pdf.

[12] S. Wasserman and J. Galaskiewicz, *Advances in Social Network Analysis: Research in the Social and Behavioral Sciences*, SAGE Publications, 1994.

[13] A. Yavari, H. Hassanpour, B. Rahimpour Cami, and M. Mahdavi, *Event prediction in social network through Twitter messages analysis*, Soc. Network Anal. Min. **12** (2022), no. 1.

[14] A. Yavari, H. Hassanpour, B. Rahimpour Cami, and M. Mahdavi, *Election prediction based on sentiment analysis using Twitter data*, Int. J. Engin. Trans. B: Appl. **35** (2022), no. 2, 372–379.

[15] A. Yavari and H. Hassanpour, *Election prediction based on messages feature analysis in Twitter social network*, Int. J. Engin. Trans. C: Aspects **36** (2023), no. 6, 1179–1184.