A Geofencing-based Recent Trends Identification from Twitter Data

IOP Conf. Series: Materials Science and Engineering **769** (2020) 012008 doi:10.1088/1757-899X/769/1/012008

A Geofencing-based Recent Trends Identification from Twitter Data

M. Saef Ullah Miah[1], M. Sadid Tahsin[2], Saiful Azad[1,3,*], Gollam Rabby[1], M. Sirajul Islam[2], Shihab Uddin[2], M. Masuduzzaman[2]

[1]Faculty of Computing, University Malaysia Pahang, 26300 Gambang, Kuantan, Pahang, Malaysia

[2]American International University-Bangladesh, Ka-66/1, Kuratoli Road, Kuril, Khilkhet, Dhaka-1229, Bangladesh

[3]IBM CoE, UMP, 26300 Gambang, Kuantan, Pahang. Malaysia

E-mail: `saifulazad@ump.edu.my (*corresponding author)`

August 2019

Abstract. For facilitating users from information overloading by finding recent trends in twitter, several techniques are proposed. However, most of these techniques need to process extensive data. Therefore, in this paper, a geofencing-based recent trends identification technique is proposed, which acquires data based on a geofence. Afterwards, they are cleaned and the weight of these tweet data is calculated. For that, the frequency of tweet texts and hashtags are taken into account along with a boosting factor. Thereafter, they are ranked to recommend recent trends to the user. This proposed technique is applied in developing a system using Java and python. It is compared with other relevant systems, where it demonstrates that the performance of the proposed system is comparable. Over and above, since the proposed system integrates geofencing feature, it is more preferable over other systems.

1. Introduction

Since microblogging social media platforms including Twitter, FriendFeed, Dailybooth, and Tumblr have become a popular communication and information search tool, the social content recommendation has risen to a new dimension [1]. Among these platforms, twitter data are preferred by most of the researchers or industries due to its large volume of users — around 321 million claimed monthly users [2] — which produce a large volume of data. Although, these data provide us with a "gold-mine" of real-world information; they lead us to an enormous problem, called information overload [3]. Due to this, the most important, interesting, and relevant tweets remain unnoticed by the users. For more details, if an active user follows 80 users on an average, s/he may receive around 1000 tweets on an average [4]. Among these, many are personal or irrelevant tweets that are unworthy of reading.

Geofencing-based Recent Trends Identification

To tackle this issue, twitter has introduced the hashtag — a word or a phrase without spaces prefixed with the hash symbol; and hence, the name. It facilitates to spreads trendy topics quickly among millions of users. Again, there remain several other important trends unidentified due to improper or no tagging. To overcome this issue, several researchers have proposed various solutions, e.g., URL recommendations for twitters, organizing trending topics in the user's timeline, recommending followers and tweets, personalized recommendation of twitter tweets [5, 6, 7, 8, 9]. However, only a few techniques are proposed that identifies trends based on geofencing — a virtual perimeter for a real-world geographic area — which are taken into account during experimental evaluations in Section 2. Again, most of these techniques do not take both hashtag and twitter text into consideration for discovering recent trends.

Therefore, in this paper, a geofencing-based recent trends identification technique is proposed utilizing twitter data. In the proposed technique, both hashtags and twitter texts are examined to identify the recent trends. The objective of this work is to assist people by offering them the most relevant and interesting trends based on geofencing; and thus, save valuable searching time for those from the ocean of data. In addition, the proposed technique also ranks the identified trends to facilitate the selection.

The rest of the subsequent sections of this paper are organized as follows. The most relevant techniques to the proposed technique are discussed in Section 2 with their known limitations and/or features to establish the necessity of proposing a new such technique. The details of the proposed technique are mentioned in Section 3. Afterwards, the performance of the proposed technique is evaluated with the relevant experiments in Section 4. This paper ends with concluding remarks in Section 5.

2. Related Works

Emotion identification employing hashtag remains an important research area to investigate. In [10], the authors endeavor to identify the emotions of different human being during crisis period. They collected tweets of various subjects after the hurricane sandy, which is considered as the deadliest, the most destructive, and the strongest hurricane of the 2012 Atlantic hurricane season. This research focuses only on analyzing the tweets of the victims to identify their emotions over the crisis. Again, in [13], another research is conducted to identify students' emotions employing twitter data. Generally, it has been observed that students cannot express their emotions to their teacher. Therefore, they prefer social media as their platform to express their emotions. In this paper, the students' tweets are acquired and data mining techniques are applied for identifying their emotions. However, the proposed technique would analyze the tweets to find the contemporary trends of a certain geographical location.

For identifying the emotions and moral relationship between them, twitter data are analyzed in [11]. Here, the authors gathered a vast amount of data from the social media. Later, they are transformed into a meaningful data. Afterwards, they are analyzed for better understanding of the audiences' emotions even before an event occurs. However,

in this paper, a vast amount of data is gathered to identify the most relevant current trends which is happening in a virtual geographical location or geofence.

Twitter data are also used to identify human behavior over the social network. In [12], a such study is performed where data from Cyworld are acquired. Afterwards, they are analyzed to identify until what extent people nowadays are involved in social media and how they behave over the social network. However, the proposed technique crawls over twitter tweets to analyze trend behavior over the social network.

In [5], a technique is proposed that can identify and recommend personalized trends to the users to aid them from searching a huge number of tweets. This proposed technique utilized term frequency and inverse document frequency (tf-idf) to identify the most important term along with collaborative filtering. However, although our proposed technique utilizes tf-idf to identify the important terms, hashtag also plays an important role in identifying the recent trends. Again, the geofencing-based identification, which is necessary for the reducing number of data as well as the processing time and focus on user interest reign, make the proposed technique distinguish over other existing technique.

3. Proposed Technique

The proposed technique is divided into three phases, namely *i*)data acquisition, *ii*) recent trend identification, and *iii*) recommending recent trends and topic-based searching. All these phases are explained in details below.

3.1. Data Acquisition

Since the proposed technique identifies the recent trends based on geofencing, data must be acquired from the Twitter accordingly. Generally, tweet data are global; therefore, it is necessary to acquire data in such a way that it enables geofencing. In this proposed technique, tweet data are acquired using tweepy library, which is a python library that can acquire real time data. Afterwards, the desired location is set before acquiring the data. All the data are acquired in json format in a . json file. Then, they are converted to .CSV (Comma Separated Value) file and stored in a MySql database.

3.2. Recent Trends Identification

The process of recent trends identification starts with fetching tweets from the database, where tweets in English is only taken into account. Afterwards, stopwords, emojis, and URLs are removed from the tweet data. Let us called this as "tweet text" to distinguish it from the tweet data.

These tweet texts are further examined to identify the hashtags since they are incorporated to facilitate users in finding messages with a specific theme or content. It is possible to identify many recent trends analyzing these hashtags as demonstrated in [1]. In the proposed technique, they also play an important role in identifying the

Geofencing-based Recent Trends Identification

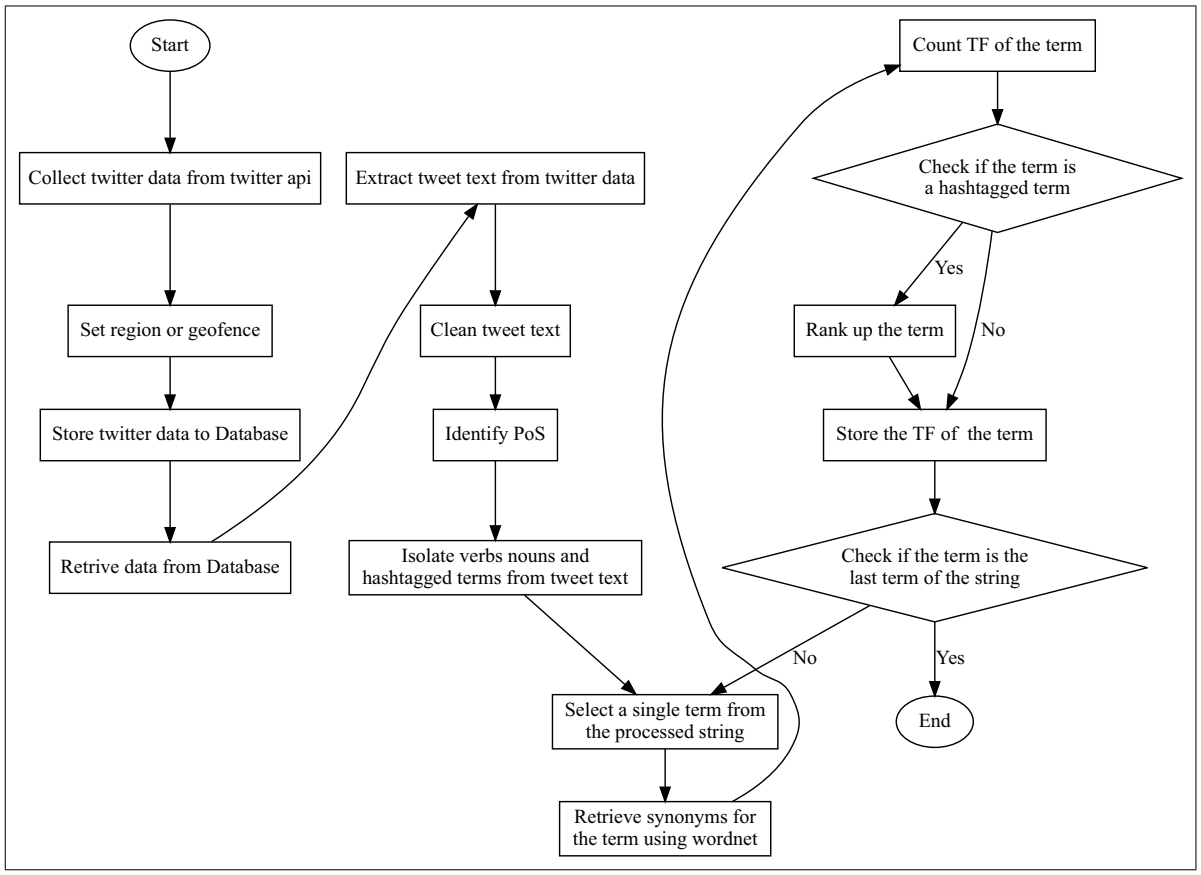


Figure 1: Steps of the proposed technique

recent trends along with other factors, which are mentioned below. Hashtags are easy to identify in a tweet text since all of them are structured and start with the "#" sign. After discovering all the hashtags from a tweet text, they are removed and saved in a list, λ_h.

Afterwards, all the nouns and verbs are identified and extracted from a tweet text; and for that, the wordnet library [17] is utilized. Later, only these nouns and verbs are considered as the representative of the entire tweet text and saved in a list, λ_s. Afterwards, the frequency of each tweet text is calculated to contribute in finding the recent trends. In the proposed technique, it is performed in the following way. Each extracted tweet text is passed to a function where a key is generated and associated with it and checked the existence of that key in a list, λ_k, which saved two-tuple data — the key and its associated count. If it is a unique key, it is appended to the list with initializing the count to 1; otherwise, the count is increased by 1. The same procedure is also followed for the hashtags in λ_h.

Then, to identify the recent trends, a matching between the hashtags and the tweeter texts are performed. Since the structure of a tweet text is different (i.e., a sentence) from that of hashtags; therefore, a substring-based matching is conducted. If

Geofencing-based Recent Trends Identification

Algorithm 1 Algorithm for finding the trends from twitter data

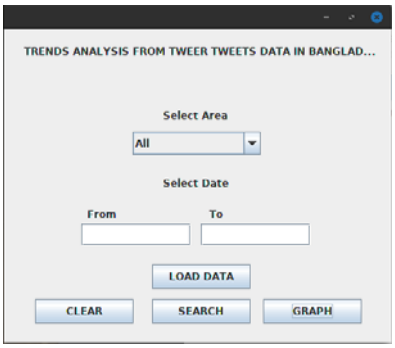
String tweets = "", hash-tags = "", all_tweets
for ($i \leftarrow 0$ to last row of the tweet table)
$str[i] \leftarrow str[i]$.removeAll(comma,dots , colon etc)
$str[i] \leftarrow str[i]$.removeAll(extra white space to single space)
$new\_str[array] \leftarrow str[i]$.split('single space')
for ($j \leftarrow 0$ to last index of new_str)
if $new\_str[j].startsWith("@http")$ or $new\_str[j].startsWith("http")$ **then**
    $new\_str[j] \leftarrow$ ""
else
    $tweets+ = new\_str[j]$
end if
if $new\_str[j].startsWith("\#")$ **then**
    $\lambda_h+ = new\_str[j]$
else
    $tweets+ = new\_str[j]$
end if
end for
$nt[array] \leftarrow tweets.split(singlespace)$
for ($m \leftarrow 0$ to last index of new_tweets)
if $nt[m]$ is 'Noun' **then**
    $nt[m] \leftarrow synonym[0].(nt[m])$
    $\lambda_s+ = nt[m]$
else
    $\lambda_s+ =$ ""
end if
if $nt[m]$ is 'Verb' **then**
    $nt[m] \leftarrow synonym[0].(nt[m])$
    $\lambda_s+ = nt[m] +$ "."
else
    $\lambda_s+ =$ ""
end if
end for
$sentence[array] \leftarrow \lambda_s.split(fullstop)$
for ($k \leftarrow 0$ to last index of sentence)
generate unique Key for similar sentence
if same unique Key **then**
    $count+ = 1$
    $\lambda_s(key, count)$
else
    $count \leftarrow 1$
    $\lambda_s(key, count)$
end if
end for
for ($k \leftarrow 0$ to last index of λ_h)
generate unique Key for similar hashtags
if same unique Key **then**
    $count+ = 1$
    $\lambda_h(sentence, count)$
else
    $count \leftarrow 1$
    $\lambda_h(hashtag, count)$
end if
end for
for ($p \leftarrow 0$ to last λ_s)
$tw[array] \leftarrow \lambda_s[p].sentence.split(singlespace)$
for ($q \leftarrow 0$ to last λ_h)
if $HTL.hashtag.matches(\lambda_s[p].sentence)$ **then**
    $\lambda_s[p].count \ * = \alpha$
end if
end for
end for
end for

a matched found, the count of that tweet text is recalculated as follows:

$$\kappa'_t = \kappa_t \times (\alpha \times \kappa_h) \tag{1}$$

where, κ_t is the previous frequency count, κ'_t is the new frequency count, κ_h is the count of the hashtag, and α is a constant, which is utilized as the boosting factor. The value of α must be more than 1. As could be observed by Equation 1 is that when a tweet text or a substring of a tweet text is matched with a hashtag, its frequency is increased at least α times of the existing count. The justification of this increment is that it has been observed that hashtags are the most influential terms in a tweet data that carry recent trend related information as demonstrated in [1]. All these steps of the proposed technique are illustrated in Algorithm 1.

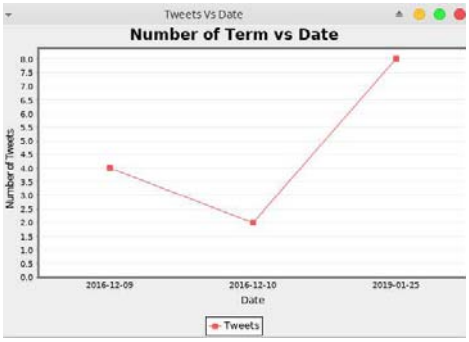
Geofencing-based Recent Trends Identification



(a) Java based GUI for the trends analysis system

(b) GUI for providing search terms

(c) Data found from database related to the search term



(d) Graphical representation of search term's occurrences

Figure 2: Four print-screens of the implemented system that is based on the geofencing-based proposed technique.

3.3. Recommending Recent Trends and Topic-based Searching

After calculating frequencies of all the tweet texts, they are ranked based on their frequency counts. Among them, only top-N trends are recommended to the users. For that, any advanced sorting technique would be adequate. However, in our case, a quick sort is utilized since its time complexity is $O(nlog(n))$.

Again, in the developed system employing the proposed technique, an option of searching any specific topic is also integrated. For that, a topic needs to be provided in the system. Afterwards, a search is initiated to match the topic with the tweets. Then it provides date wise count of that specific word.

4. Experimental Evaluation

A system has been implemented following Algorithm 1. Using the system, a user can search a topic or all the recent trends between a date or any term related within an epoch can be visualized as demonstrated in Figure 2. The system was developed in Java and python for the cross platform advantage; thus, this system can be employed in any platform. Once the data is loaded into the system, it takes a minimal time to

Table 1: Comparison between Three popular system and our system

	keyhole	**tweetreach**	**tweetbinder**	**proposed system**
Search Type	hash tag based	hash tag and keyword based	hash tag and keyword based	hash tag, keywords and geo location based
Licensing	paid and free with minimal feature	paid and free with minimal feature	paid and free with minimal feature	will be open-sourced after publication of the paper
Accuracy	90% on free version and stored data	95% on free version and stored data	95% on free version and stored data	95% on stored data with geo location, hashtags and keywords

identify and visualize the data.

To find out the effectiveness of the developed system, it is compared with other three most relevant existing systems, namely keyhole [14], tweetreach [15], and tweetbinder [16]. Here, it is noteworthy to mention that all these systems are freemium systems, i.e., some features of these systems are free and some of them are premium.

For our experimental evaluation, a location has been chosen for acquiring data for a period of time where some important events have taken place. In this paper, Bangladesh is chosen as the geofencing location. Again, we have tested different keywords in the free versions of the aforementioned. The results are noted in Table 1. As it could be observed from the table is that tweetreach, tweetbinder, and the proposed system demonstrate comparable performance, i.e., 95% accuracy, which is well over keyhole system with 90% accuracy. However, since the proposed system integrates geofencing for finding the recent trends, it requires only limited data to store and process; and thus, preferable over other techniques. Full system implementation can be found at the following repository hosted on github.com. Repository link: https://github.com/ping543f/A-Geofencing-based-Recent-Trends-Identification-from-Twitter-Data.git

5. Conclusions

This paper presents a geofencing-based recent trends identification technique using twitter data. In the proposed technique, data are acquired based on the geofence. Afterwards, they are cleaned and the weight of each tweet text is calculated based on the frequency of itself and substring matching of hashtags. Then, they are ranked

to recommend to the user. The proposed technique is applied to developed a system in Java and python and is compared with other relevant systems. The experimental results demonstrate that the performance of the proposed system is comparable with other compared systems. However, due to its geofencing feature, it is more preferable over other systems.

Acknowledgement

This work has been partially supported by the RDU grant, RDU180359, of University Malaysia Pahang, Malaysia.

Referencing

[1] Eriko Otsuka, Scott A. Wallace, and David Chiu. *A hashtag recommendation system for twitter data streams*. Computational Social Networks, 3(3), 2016.

[2] Hamza Shaban. *Twitter reveals its daily active user numbers for the first time* The washington Post, 2019.
`https://www.washingtonpost.com/technology/2019/02/07/twitter-reveals-its-daily-active-user-numbers-first-time/?noredirect=on&utm_term=.778eba392257`

[3] C. Borgs, J. Chayes, B. Karrer, B. Meeder, R. Ravi, R. Regans, and A. Sayedi. *Game-theoretic models of information overload in social networks*. Algorithms and models for the Web-Graph, Springer, 146-161, 2010.

[4] Z. Qu and Y. Liu. *Interactive group suggesting for twitter*. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers, 2:519523, 2011.

[5] Vineeth Rakesh, Dilpreet Singh, Bhanukiran Vinzamuri, Chandan K. Reddy. *Personalized Recommendation of Twitter lists Using Content and Network Information* In Proceedings of the Eight International AAAI Conference on Weblogs and Social Media, 2014.

[6] F. Abel, Q. Gao, G.-J. Houben, and K. Tao. *Analyzing temporal dynamics in twitter profiles for personalized recommendations in the social web*. In Proceedings of the 3rd International Web Science Conference, ACM, 2011.

[7] M. G. Armentano, D. Godoy, and A. Amandi. *Topology-based recommendation of users in microblogging communities*. Journal of Computer Science and Technology, 27(3):624634, 2012.

[8] M. S. Bernstein, B. Suh, L. Hong, J. Chen, S. Kairam, and E. H. Chi. , E. H. 2010. *Eddi: interactive topic-based browsing of social status streams*. In Proceedings of the 23rd annual ACM symposium on User interface software and technology, ACM, 303312, 2010.

[9] J. Hannon, M. Bennett, and B. Smyth. *Recommending twitter users to follow using content and collaborative filtering approaches*. In Proceedings of the fourth ACM conference on Recommender systems, ACM, 199206, 2010.

[10] J. Brynielsson, F. Johansson, C. Jonsson, and A. Westling. *Emotion classification of social media posts for estimating peoples reactions to communicated alert messages during crises*. Security Informatics, 2014.

[11] D. M. Haughton, J. J. Xu, D. J. Yates, and X. Yan. *Introduction to Data Analytics and Data Mining for Social Media Minitrack*. Hawaii International Conference on System Sciences, 2016.

[12] Y.-Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong. *Analysis of topological characteristics of huge online social networking services*. In Proceedings of the 16th international conference on World Wide Web. ACM, 2007.

[13] X. Chen, M. Vorvoreanu, and K. Madhavan. *Mining social media data for understand-ing students learning experiences*. IEEE Transactions on Learning Technologies, 7(3): 246-259, 2014.

[14] *Whatever happens in the world shows up on Twitter*, 2019.
texttthttps://keyhole.co/twitter-analytics/
[15] *Free Twitter Analytics Report from Union Metrics*, 2019.
`https://tweetreach.com/`
[16] *Analyze, classify, and display twitter and Instagram content*, 2019.
`https://www.tweetbinder.com/`
[17] George A. Miller. *WordNet: A Lexical Database for English.* Communications of the ACM, 38(11):39-41, 1995.
[18] *Hale, S. A. (2014) Global Connectivity and Multilinguals in the Twitter Network. In Proceedings of the 2014 ACM Annual Conference on Human Factors in Computing Systems, ACM (Montreal, Canada).*