2020 IEEE Region 10 Symposium (TENSYMP), 5-7 June 2020, Dhaka, Bangladesh

Discovering and Tracking Query Oriented Topical Clusters in Online Social Networks

Tanjim Taharat Aurpa*, Fatema Khan† and Md Musfique Anwar‡
Department of Compurter Science and Engineering, Jahangirnagar University,Savar,Bangladesh
Email: (taurpa22 *, fatemakhankgsc†)@gmail.com, manwar@juniv.edu‡

Abstract—**Online Social Networks (OSNs) are comprehensive media that help individuals to be connected through social networking sites (SNS) such as Twitter, Instagram, etc. People share their interests, activities and can exchange ideas. OSNs are typically large in size and complex as those media have an enormous number of users and multi kind relationships among them. Users reveal their interests in diverse topics in OSN and mostly, users' degree of topical interest changes over time. Tracking users' interests from such SNS and grouping users having similar interests based on that becomes significant for various domains. In this paper, we pay attention to identify and track users' topical interests on Twitter over time. Next, we group users with similar degrees of interest in different clusters. We perform experiments on real datasets and got interesting results.**

Keywords–Online Social Network, Active user, Topical Clusters

I. Introduction

According to Social Media Statistics in 2019, the number of social media users are 3.2 billion where 330 million users are from Twitter [1]. Twitter is a micro-blogging website on which users post micro messages known as tweets. Several former studies endeavor to infer topics from Twitter by the contents of the user's tweets. Although inferred topics are not significant sometimes as tweets often comprise contents about users' routine activities. But users' topical interests vary over time. Again, users on OSNs (Twitter) are often clustered. Many algorithms are existed for clustering users according to their similar topical interests. Consequently, these approaches unable to trace active clusters. But it is necessary to consider user activeness along with topical information for cluster detection, particularly for advertising and marketing intentions. In our proposed methodology, for a given input query consisting of one or more topics, we identify and track users' topical interests as well as cluster similar users based on their topical degree of interests over time. In summary, our contributions are as follows:

We proposed an approach to identify and track users' topical degree of interests over time and showed how users' activeness is changing concerning time. We accomplished comprehensive experiments using real data sets to show that how much efficient our approach.

II. Problem Description and Proposed Framework

First, we formally formulate the problem of detecting the active local community in the social network. Then we give an overview of our proposed framework.

A. Problem Description

Before defining the problem definition, we introduce several related concepts.

Attributed Graph: An attributed graph is denoted as $G = (U, E, \mathcal{T})$, where U is the set of nodes (users), E is the set of links between the nodes (connections or virtual social relationships between users in G, such as the *following* relationships in Twitter), and $\mathcal{T} = \{T_1, T_2..., T_m\}$ is the set of topics discussed by the social users U.

Topic: A topic is a distribution over words, i.e., it contains most representative words for that topic. For example, *politics* topic has words like election, government, democratic, parliament, etc. about politics.

Activity: Activity refers to an action that a user performs at a time point. For example, a user u in Twitter posts a tweet (message) containing a specific topic T_i at time t_j. This activity is recorded as an activity tuple $\langle u, T_i, t_j \rangle$. An activity stream S is a continuous and temporal sequence of activities i.e. $S = \{s_1, s_2, ..., s_r, ..\}$ such that each object (s_i) corresponds to an activity tuple.

Query: An input query $Q = \{\mathcal{T}_q\}$ consisting a set of query Topics $\mathcal{T}_q = \{T_i, T_{i+1}..., T_n\}$.

Sliding Time Window: A window of a predefined length len is moved over the activity stream S and specifies the intervals to analyze. Let $\Gamma =< t_1, t_2, ..., t_n >$ be a sequence of points in time, I_m an interval $[t_{i-len}, t_i]$ of len, where $0 < len \leq i$. We partition Γ into set of equal-length intervals denoted as $\mathcal{I} = \{I_1, ..., I_m\}$. We consider an *overlapping window* partially overlaps with the prior window. The degree of overlap is controlled by the parameter Δt.

Topical Interest Score: For each active user $u_i \in U$, we compute her topical interest score (denoted by Ω) to measure the involvement of u_i towards the given query attributes $\mathcal{T}_q$ of Q, using Equations 1 and 2, where $\psi_{u_i} \in Q$.

$$\Omega_{I_k}(u_i, \psi_{u_i}) = \frac{|ACTS(u_i, \psi_{u_i})|}{\lambda_{(Q, U^Q_{I_k})}} \tag{1}$$

where, $ACTS(u_i, \psi_{u_i})$ indicates the set of activities containing the set of topics $\psi_{u_i} \subseteq Q$ performed by u_i and $\lambda_{(Q, U^Q_{I_k})}$ denotes the *average* number of activities related to Q performed by $U^Q_{I_k}$ in G where $U^Q_{I_k}$ indicates only those users who posted tweets related to Q at time interval I_k.

$$\lambda_{(Q, U^Q_{I_k})} = \frac{\sum_{u_i \in U^Q_{I_k}} |ACTS(u_i, \psi_{u_i})|}{|U^Q_{I_k}|} \tag{2}$$

978-1-7281-7366-5/20/$31.00 © 2020 IEEE

Then, the activeness (denoted as σ) of u related to Q is

$$\sigma_{(u_i,\psi_{u_i})} = \frac{\Omega_{I_k}(u_i,\psi_{u_i})}{max_{u_z \in U^Q_{I_k}}\{\Omega_{I_k}(u_z,\psi_{u_z})\}} \tag{3}$$

Problem Definition: Given an attributed graph *G = (U, E, $\mathcal{T}$)*, an input query $Q = \{\mathcal{T}_q\}$, we want to group users into three different clusters (namely $\mathcal{C_H}$, $\mathcal{C_M}$ and $\mathcal{C_L}$ as high, medium and low active groups respectively) based on their topical interest scores.

B. Example

We consider an attributed graph denoted as *G = (U, E, $\mathcal{T}$)*, where a set of users *U=(u_1,u_2,u_3,.....,u_{16})* and the set of connections among them is *E*. We shift the time window for a fixed topic $\mathcal{T}$. Let's assume all users of *G* have topical interest score measured by our proposed methodology as shown in Table. I. In time window I_1, we get three clusters mentioned in Fig. 1. Here high active users, medium active users and low active users are clustered as C_H, C_M and C_L respectively. In I_1 we can observe that C_L has 3 members, C_M has 5 members and C_H has 7 members. These clusters' size are detected using the score in Table I. In Fig. 1, we use different colors for different clusters' members.

Next, we shift the time window from I_1 to I_2 and observe these clusters. The size of cluster C_H, C_M, C_L has been changed and become 4, 5, and 6 respectively. The changes started to be visible by mixed colors in a cluster. Though size doesn't change for cluster C_M, still changes occur in its members which is clearly understood by the users' mixed colors in the cluster.

Here users have changed their clusters as well as their topical interest score. The score of user u_1, u_2, u_5, u_6 has been decreased to the lower cluster. User u_{11} is dropped from the clusters and becomes inactive.

On the other hand, user u_3, u_4, u_7, u_{10}, u_{13}, u_{15} have scored higher than previous time window and promoted to higher one. User u_{16} has added to C_L who was inactive before. Remaining users may have different scores than previous but this difference doesn't visualize any changes on cluster's membership.

In our proposed methodology we have tracked cluster in another way by fixing the time window and shifting the value of query topic $\mathcal{T}$. In that case, clusters changes are observed in the same way of this given example.

C. Overview of the Proposed Framework

An attributed social graph is essentially a graph associated with text strings or keywords. E.g., Twitter users post tweets on different topics such as politics, business, and marketing, etc [3]. Sometimes, the topics in the tweets are not mentioned explicitly, for example, the topic of a tweet on Twitter is mentioned via *hashtags*. As a result, we need to properly identify the topic associated with each tweet.

Our proposed approach has two steps for clustering users according to Q. Firstly, we use the topic modeling method Twitter Latent Dirichlet Allocation (T-LDA) [2] on tweets to identify the latent topics. Next, we apply our proposed algorithm to find desired clusters on different time windows.

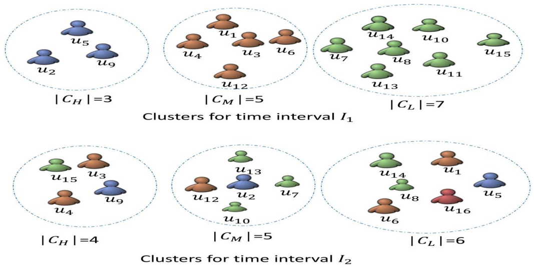

Fig. 1: Changes in clusters in different time window for query $Q = \{\mathcal{T}_1\}$

Table I: Users Topical Interest Score

u_i	I_1	I_2	u_i	I_1	I_2
u_1	0.53	0.39	u_9	0.87	0.91
u_2	0.8	0.45	u_{10}	0.25	0.65
u_3	0.6	0.78	u_{11}	0.3	Drop
u_4	0.58	0.91	u_{12}	0.59	0.61
u_5	0.85	0.31	u_{13}	0.22	0.66
u_6	0.63	0.25	u_{14}	0.25	0.2
u_7	0.28	0.55	u_{15}	0.3	0.82
u_8	0.27	0.33	u_{16}	Inactive	0.35

III. Topical Cluster Detection and Tracking

We apply the proposed method in a large Twitter dataset.

A. Topic Detection from Tweets

Topic modeling is usually done by detecting patterns in a collection of documents called a *corpus*, and grouping the words used into topics. In Twitter, the topics are not explicitly expressed in tweets. So, we choose Twitter Latent Dirichlet Allocation (T-LDA), a variation of LDA, to infer the latent topics from the tweets.

We select tweets of 2000 users posted tweets from June 11, 2009, to June 30, 2009, and apply T-LDA to infer 100 topics from those tweets. Each tweet is indicated by a topic number in T-LDA. Each topic contains a set of related words. T-LDA is based on the assumptions given below.

There are *k* topics in Twitter. Distribution over *k* topics represents that each user has own topic interests *u*. Topic *k* is assigned to each *u* depending on the topic interests ϕ_u Each word in the tweet assigned by topic *k* is generated from a topic word distribution θ_k. The latent value *y* determines whether the word is a background word or a topic word.

This generative process is graphically represented in Fig. 2. Table II shows sample word topic distribution.

B. Topical Clusters Detection Algorithm

We develop an algorithmic framework to detect and track users' topical clusters.

Algorithm overview. The algorithm, called `Query Algorithm`, finds the set of users $U^Q_{I_k}$ from U for a given Q at each time interval I_k at first and then computes users' interest score $\sigma_{(u_i,\psi_{u_i})}$ (line 1-4). Finally, it groups active users into different clusters based on users' topical interest scores (line 5-10). It outputs Φ^Q at eachI_k (line 11).

IV. Experimental Evaluation

We conduct our experiments on a real Twitter dataset which are performed on an Intel(R) Core(TM) i5-7220U 2.5 GHz Windows 10 PC with 8 GB RAM.

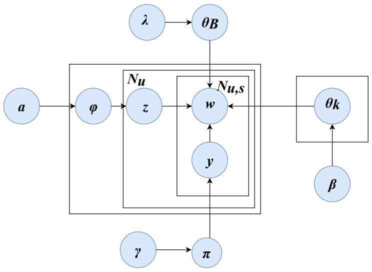

Fig. 2: Graphical Representation of Twitter-LDA Model

Table II: Sample Word Topic Distribution in T-LDA Model

Topic 1	**Topic 2**	**Topic 3**	**Topic 4**
Iran	film	died	twitter
democracy	terminator	sad	message
protest	media	michael	followers
election	screen	jackson	activated

Algorithm 1 `Query Algorithm`

```
Input: $G = (U, E, \mathcal{T}), \mathcal{I}, Q, \theta$
Output: set of topical clusters $\Phi_Q = \{\mathcal{C}_\mathcal{H}, \mathcal{C}_\mathcal{M}, \mathcal{C}_\mathcal{L}\}$
for each $I_k \in \mathcal{I}$ do
    select $U_{I_k}^Q$ from $U$    ▷ each $u_i \in U$ has to perform certain number of actions related to $Q$
    for each $u_i \in U_{I_k}^Q$ do
        compute $\sigma_{(u_i, \psi_{u_i})}$
        if $\sigma_{(u_i, \psi_{u_i})} \geq 0.75$ then
            $\mathcal{C}_\mathcal{H}.add(u_i)$
        else if ($\sigma_{(u_i, \psi_{u_i})} \geq 0.4$ and $< 0.7$) then
            $\mathcal{C}_\mathcal{M}.add(u_i)$
        else if ($\sigma_{(u_i, \psi_{u_i})} \geq 0.25$ and $< 0.4$) then
            $\mathcal{C}_\mathcal{L}.add(u_i)$
Output the set of topical clusters $\Phi_Q$ at each time interval $I_k$
```

A. *Data set*

We conduct our experiment on a Twitter dataset named SNAP[1]. SNAP contains 467 million Twitter posts from 20 million users from June 1, 2009 to December 31, 2009. We randomly choose 2,000 users and consider their tweets from June 11, 2009 to June 30, 2009.

B. *Experimental Results*

In SNAP dataset, we consider users' tweets for query Q consists of *Iran Election*, *Entertainment* and both. We choose the length of each time window (I_k) as 7 days and then shift each I_k by 2 days i.e. $\Delta t = 2$. At each I_k, we cluster the *active users* (users who having at-least 5 activities related to Q at I_k) into *high* ($\mathcal{C}_\mathcal{H}$), *medium* ($\mathcal{C}_\mathcal{M}$) and *low* ($\mathcal{C}_\mathcal{L}$) topical clusters on the basis of their interest score. Here we indicate *high* ($\mathcal{C}_\mathcal{H}$) with interest score greater than 0.75, *medium* ($\mathcal{C}_\mathcal{M}$) with interest score between 0.40 and 0.75 and *low* ($\mathcal{C}_\mathcal{L}$) with interest score between 0.25 and 0.40 . We observe the changes in the clusters by varying topic and time window.

In our experiment, we find that 947 users posted on *Iran Election* from June 11, 2009 to June 23, 2009. Table III shows the differences in clusters' sizes as well as the number of new (or existing) members added (or dropped) at different time windows (len(I_k) = 7 days, Δt = 2 days).

Table III: Tracking Clusters at different I_k for topic *Iran Election*

$I_k \longrightarrow$	(11/06-17/06)	(13/06-19/06)			(15/06-21/06)		
$\mathcal{C}\downarrow$	C	C	Drop	Add	C	Drop	Add
$\mathcal{C}_\mathcal{H}$	9	12	1	4	6	7	1
$\mathcal{C}_\mathcal{M}$	16	15	7	6	10	8	3
$\mathcal{C}_\mathcal{L}$	23	33	12	22	25	12	4

We observe the following changes in time window (13/06 - 19/06):

- 1 member is dropped to $\mathcal{C}_\mathcal{M}$ from $\mathcal{C}_\mathcal{H}$. Again 4 members from $\mathcal{C}_\mathcal{M}$ are added to $\mathcal{C}_\mathcal{H}$. 2 members from $\mathcal{C}_\mathcal{M}$ become inactive user and 1 is dropped to $\mathcal{C}_\mathcal{H}$, 4 members to $\mathcal{C}_\mathcal{H}$. Again, 4 members from $\mathcal{C}_\mathcal{L}$ are added to $\mathcal{C}_\mathcal{M}$, 1 member from $\mathcal{C}_\mathcal{H}$ and new user become active user who enter into $\mathcal{C}_\mathcal{M}$ directly. 5 members from $\mathcal{C}_\mathcal{L}$ dropped to $\mathcal{C}_\mathcal{M}$ and 7 are became inactive user. 1 member from $\mathcal{C}_\mathcal{M}$ is added to $\mathcal{C}_\mathcal{L}$ and 21 members become active user and enter into $\mathcal{C}_\mathcal{L}$

We again shift the time window for two days and observe users' activities from June 15, 2009 to June 21, 2009 and be able to indicate the following changes:

- From $\mathcal{C}_\mathcal{H}$, one member is dropped to $\mathcal{C}_\mathcal{M}$, 3 members are dropped to $\mathcal{C}_\mathcal{L}$ and 3 other members become inactive user. From $\mathcal{C}_\mathcal{M}$, 6 members become inactive user and 1 member is added to $\mathcal{C}_\mathcal{L}$ and another 1 member to $\mathcal{C}_\mathcal{H}$. 1 member from $\mathcal{C}_\mathcal{L}$ is added to $\mathcal{C}_\mathcal{M}$, 2 users became active who enter into $\mathcal{C}_\mathcal{L}$ and 11 members become inactive user.

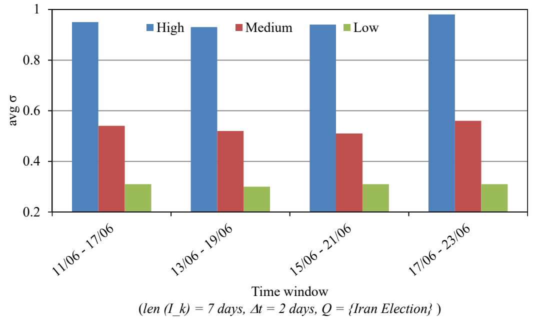

Fig. 3: Average interest score at different time intervals at each topical cluster for topic *Iran Election*

Fig. 3 shows the average interest scores of the members of each cluster at different time intervals for the topic *Iran Election*. We see that the members in High and Medium clusters have more average score at time interval from 17/06/2009 to 23/06/2009.

Next, we trace another observation by changing the the value of Q=*Iran Election*, *Entertainment*. Considering the clusters' sizes for different topics at different time window, we can detect bursty topics.

C. *Real-time Bursty Topic Detection by Tracking Clusters' Size*

Bursty topic *is a behavior associated to a topic within a time interval in which it has been extensively treated but rarely before and after* [7]. We present a method by tracking the topical clusters to detect bursty topic. For bursty topic

[1] http://snap.stanford.edu/data/twitter7.html

Table IV: Tracking cluster sizes Real-time Bursty Topic detection

$Q \downarrow$	$I_k \longrightarrow$	11/6-15/06	14/6-18/06	17/6-21/06	20/6-24/06	23/6-27/06	26/6-30/06
Iran Election	$\lvert\mathcal{C}_{\mathcal{H}}\rvert$	7	11	3	1	4	2
	$\lvert\mathcal{C}_{\mathcal{M}}\rvert$	17	13	7	5	8	0
	$\lvert\mathcal{C}_{\mathcal{L}}\rvert$	21	27	17	7	11	6
	$Total$	45	51	27	13	23	8
Entertainment	$\lvert\mathcal{C}_{\mathcal{H}}\rvert$	3	1	2	2	2	2
	$\lvert\mathcal{C}_{\mathcal{M}}\rvert$	4	2	6	3	4	3
	$\lvert\mathcal{C}_{\mathcal{L}}\rvert$	4	1	1	4	10	4
	$Total$	11	4	9	9	16	9

the cluster size is usually low. In a certain time window, it becomes abnormally high and at next time window again the size decreases to low. On the other hand a normal topic's cluster size remains almost same for different time windows.

In Table IV, we consider two value of topic $Q = IranElection, Entertainment$ and track the size of clusters. Here we consider time window len(I_k) = 5 days, Δt = 3 days. For $Q = IranElection$ we can observe that the cluster size $|C|$ is abnormally increasing from 1st time window 11/06-15/06 to 2nd time window 14/6-18/06. As Iran Election was held on 12 June 2009, people posted the most at this time. But after June 18 the cluster sizes start to decrease and at last time window become very low. So we can say this topic drew bursts of interest in our 1st and 2nd time window and almost vanished at 6th time window. So $Q = IranElection$ is a bursty topic. For $Q = Entertainment$ the clusters' sizes never increment abnormally, again never vanish at some point. Here $|C|$ always remains near to the average value. It never draw bursts of interest. Clearly $Q = Entertainment$ is not a bursty topic. For more clarification we can observe the graph i dsrawn in Fig4. There are two lines indicating two values of Q. The red line indicates the bursty topic *Iran Election* and the blue line indicates a topic *Entertainment* which has not drawn to burst.

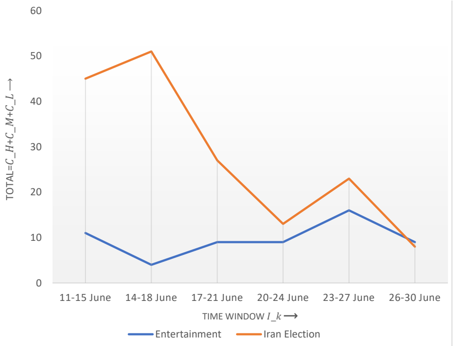

Fig. 4: Real-time Bursty Topic (*Iran Election*) detection using cluster size

V. Discussion

In this paper, we trace the changes of users' interest according to the changes in time and topics. We represent three clusters ($\mathcal{C}_{\mathcal{H}}$,$\mathcal{C}_{\mathcal{M}}$,$\mathcal{C}_{\mathcal{L}}$) to determine how their activeness or topical interest scores's change.

A. Tracking changes in cluster for one topic

We fix the value of *Q=Iran Election* and increment the time window $\mathcal{I}_{\mathcal{K}}$ by Δt = 2 days. The clusters' sizes change by this shifting. Members are dropped and added in each cluster ($\mathcal{C}_{\mathcal{H}}$,$\mathcal{C}_{\mathcal{M}}$,$\mathcal{C}_{\mathcal{L}}$). Not only new users are becoming interested but also some old users are loosing interest. So we came to this conclusion that users' interest changes by the flow of time for a constant topic.

B. Tracking changes in cluster for multiple topic

Now we consider *Q=Iran Election, Entertainment* and increment the time $\mathcal{I}_{\mathcal{K}}$ window by Δt = 3 days. We track two types of behaviour for these two values of Q. Here we see that some topic draws to burst of interest at a certain time and again these may vanish to low interest.This topics are called **Bursty topics**. We observe another type of topic which never draws to burst of interest. But near to average number of users always tweet on these topics.

VI. Related Work

Many approaches have been taken to cluster Twitter users based on diverse parameters. Michelson et al. [4] presented the outcome on inventing Twitter users' topics of interest by investigating the entities users allusion in their tweets. Liang et al. [6] proposed two collapsed Gibbs sampling algorithms to collaboratively inferring users' dynamic interests for their clustering. Another direction is to explore the content of the interactions among social users, e.g., [5], [7] to improve the quality of discovered clusters. However, the common aspect of the above works is that they did not pay much attention to users' topical activeness. Consequently, the discovered clusters cannot capture the active relationships of social users when an emerging news or event occurs in OSNs.

VII. Conclusion

Clustering analogous Twitter users based on their topical degree of interests over time is the main theme of our work. First, we have enumerated user similarities from an input query comprising of one or more topics and then grouped them according to their topical activeness. Experimental results show that our method can effectively detect and track cluster of twitter users considering users' activeness over time. In future work, we will apply our proposed method to other OSNs like Flickr, Instagram, DBLP (Co-author dataset) etc.

References

[1] Y. Lin. "10 Twitter Statistics Every Marketer Should Know in 2020 [Infographic]." November 30, 2019. [Online]. Available: Oberlo, https://www.oberlo.com/blog/twitter-statistics

[2] W. X. Zhao, J. Jiang, J. Weng, J. He, E. Lim, H. Yan, and X. Li, "Comparing twitter and traditional media using topic models", Proc. ECIR, pp. 338–349, 2011.

[3] X. Huang, and L. V.S. Lakshmanan, "Attribute-driven community search", Proc. VLDB Endowment, vol. 10, no. 9, pp. 949-960, 2017.

[4] M. Michelson and S. A. Macskassy, "Discovering Users' Topics of Interest on Twitter: A First Look", Proc. Fourth Workshop on Analytics for Noisy Unstructured Text Data (CIKM), pp. 73-80, 2010.

[5] G. Qi, C. C. Aggarwal, and T. Huang, "Community detection with edge content in social media networks", Proc. ICDE, pp. 534–545, 2012.

[6] S. Liang, E. Yilmaz, E. Kanoulas, "Collaboratively Tracking Interests for User Clustering in Streams of Short Texts", TKDE, vol. 31, no. 2, pp. 257–272, 2019.

[7] T. Yang, R. Jin, Y. Chi., and S. Zhu, "Combining link and content for community detection: a discriminative approach", Proc. KDD, pp. 927–936, 2009.