2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

Online Topical Clusters Detection for Top-*k* Trending Topics in Twitter

Md Shoaib Ahmed
Computer Science and Engr. Dept.
Jahangirnagar University
Dhaka, Bangladesh
shoaibmehrab011@gmail.com

Tanjim Taharat Aurpa
Computer Science and Engr. Dept.
Jahangirnagar University
Dhaka, Bangladesh
taurpa22@gmail.com

Md Musfique Anwar
Computer Science and Engr. Dept.
Jahangirnagar University
Dhaka, Bangladesh
manwar@juniv.edu

***Abstract*—This paper tackles the problem of detecting temporal query oriented topical clusters for top-*k* trending topics from Twitter. There is an increasing demand to identify and cluster set of users who have similar topical interests as well as certain level of *activeness* on those topics. Most existing approaches focus on the contents generated by the social users and link structure of the underlying social network. However, the degree of users' topical activeness has not been thoroughly studied to identify its effect on the formation of topical clusters. This research investigates on how the users' behaviors and topical activeness vary with time and how these parameters can be employed in order to improve the quality of the detected topical clusters for top-*k* trending topics at different time intervals. The effectiveness of our proposed activity biased weight methodology is justified using a benchmark Twitter dataset.**

***Keywords*–**Active user, Topical clusters, Trending topics

I. Introduction

Online social networks (OSNs) have gained huge popularity as social users can easily make connections with others and can share different contents (such as tweets, images, videos etc.) with any number of peers. Different people have different interests, choices, preferences and thus it is important to group similar users into common clusters. Discovering meaningful topical clusters in OSNs has recently occupied an overwhelming research interest owing to its diverse applications including online marketing, link prediction, information diffusion, friend/news recommendations etc.

A fair amount of topic-oriented methodologies have been proposed that consider the attributes of the users jointly with social connections to discover meaningful topical clusters [4], [7]. All these foregoing investigations ignore an important aspect, namely topical activeness of the social users towards query topics. As a result, the resulting clusters may contain mix of high and low active users as well as may have users who have no inclination towards the query attributes. However, we are engrossed in searching topical clusters in which users continuously pay active attention to the query attributes in a given time-period.

This paper introduces a novel concept of users' *activeness* which indicates users' topical degree of interest for a certain period of time. Our observation is that users have different degrees of topical activeness which vary widely over time. The proposed approach is commenced on measuring the degree of activeness for each candidate community member with respect to the given query attributes to enhance the quality of the detected topical clusters. Instead of giving the query topics manually, our system identifies the top-*k* trending topics by taking into account the number of mentions on that topics as well as the coverage of that topics in OSN. The main contributions of this research are summarized as follows:

- Propose a model to list top-*k* trending topics in Twitter at each time interval.
- Modeling and evaluating users' degree of activeness towards different topics of a given query.
- Conduct extensive experiments to justify the efficacy of our proposed approach using a benchmark data set.

II. Related Work

Recent 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, none of these methods address the user's degree of interest towards the given query topics as well as don't contemplate how the users' interests for the given query attributes changes with time.

III. Problem Statement

We introduce some relevant concepts before defining the problem statement.

Attributed Graph: An attributed graph is expressed as $G = (U, E, \mathcal{T})$, where U is the set of nodes (users), E is the set of links between the users, 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 collection of the 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

IEEE/ACM ASONAM 2020, December 7-10, 2020
978-1-7281-1056-1/20/$31.00 © 2020 IEEE

tweet (message) containing a specific topic T_i at time t_j. This activity is recorded as an activity tuple $\langle u, T_i, t_j \rangle$.

Sliding Time Window: 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 $\mathcal{I} = \{I_1, ..., I_m\}$.

Query: An input query $Q = \{\mathcal{T}_q\}$ consisting top-k trending Topics $\mathcal{T}_q = \{T_i, T_{i+1}..., T_k\}$ at a particular time interval.

Topical Interest Score: For each 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 $\psi_{u_i} \in Q$.

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

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 $\lambda_{(Q, U^Q_{I_m})}$ denotes the *average* number of activities related to Q performed by $U^Q_{I_m}$ in G where $U^Q_{I_m}$ indicates only those users who posted tweets related to Q at time interval I_m.

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

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

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

Problem Definition: Given an attributed graph *G = (U, E, $\mathcal{T}$)*, an input query $Q = \{\mathcal{T}_q\}$, a positive integer k, and a threshold value of θ, we want to group users into three different clusters (namely $\mathcal{C}_\mathcal{H}$, $\mathcal{C}_\mathcal{M}$ and $\mathcal{C}_\mathcal{L}$ as high, medium and low active groups respectively) based on their topical interest scores. We consider a threshold $\theta \in [0, 1]$ so that each user has to show her inclination to at least $|Q| \times \theta$ (We use the ceiling function for floating values.) topics. The low value of θ makes the query cohesiveness in relax mode while higher value of θ will impose strict cohesiveness related to Q.

IV. Topical Cluster Detection Approach

Our proposed approach has three stages as presented in Fig 1. Below, we briefly describe each of the stage.

A. Data Pre-processing for Topic Detection

In general, tweets are informally written and often contain grammatically incorrect sentence structures with misspellings and non-standard words (e.g., toook for took, goooood for good, 4eva for forever, 2day for today) etc. We performed normalization of the tweets through direct substitution of lexical variants with their standard forms with a normalization lexicon proposed by Han et al. [3].

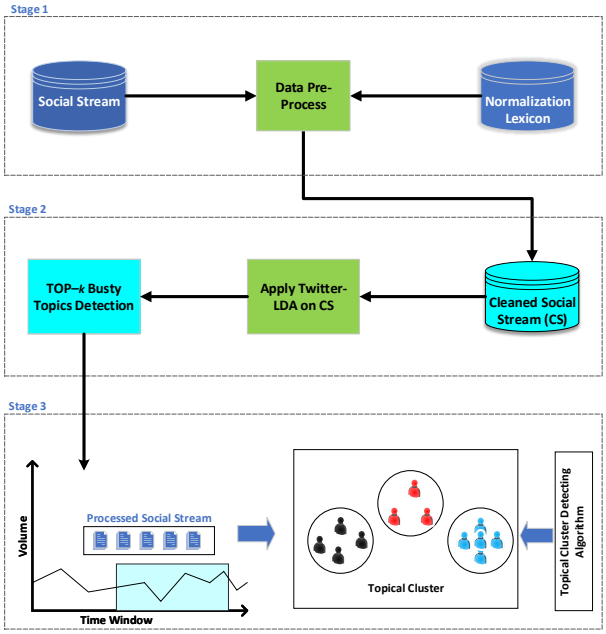

Fig. 1. Workflow of proposed framework

B. Topic Detection from Social Stream

Twitter users often use hashtags (for example, #Obama, #Ronaldo etc.) to indicate topics of the tweets. Use of hashtag is optional and there is no specific standard rules of using hashtag to mention a topic. As a result, it is difficult to correctly extract topics from hashtags and even sometimes many important topics can be skipped. So, we apply Twitter Latent Dirichlet Allocation (T-LDA) [1] for topic modeling to infer the latent topics from the tweets.

The graphical representation of T-LDA is shown in Figure. 2. The formulation of T-LDA is given below:

- Every user's topical interest ϕ_i is represented by a distribution over N topics.
- Each word is implied by topic N is analyzed from a background word distribution and topic word distribution represented by θ_B and θ_K respectively. π is Bernoulli distribution that controls the possibility between background and topic words.
- Dirichlet distributions like α, β, γ and λ govern the other multinomial distributions.
- The latent value y determines whether the word is a background word or a topic word.
- z represents the topics of words here.

C. Top-k Trending Topics from Social Stream

In our proposed model, we set the value of the query Q at each time interval I_m as the top-k trending (busty) topics at that I_m. We define trending score ($\eta_{(T_j, I_m)}$) for each topic T_j according to equation 4:

$$\eta_{(T_j, I_m)} = \alpha \times \left|ACTS(*, T_j)\right| + (1 - \alpha) \times U_{T_j, I_m} \tag{4}$$

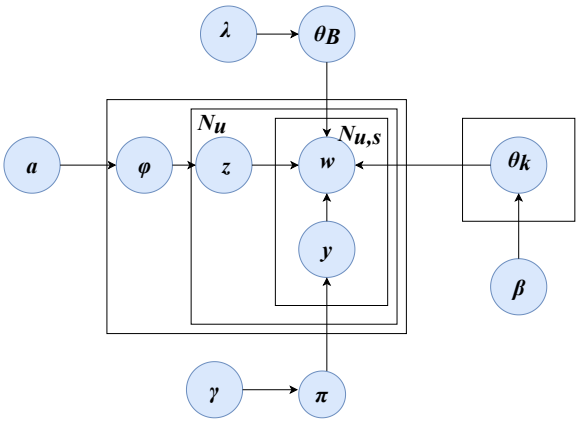

Fig. 2. Graphical Representation of Twitter-LDA Model

Algorithm 1 `Query Algorithm`

Require: $G = (U, E, \mathcal{T}), \mathcal{I}, Q, S, k, \theta, \alpha$
Ensure: set of topical clusters $\Phi_Q = \{\mathcal{C}_\mathcal{H}, \mathcal{C}_\mathcal{M}, \mathcal{C}_\mathcal{L}\}$
1: **for** each $I_m \in \mathcal{I}$ **do**
2: $Q \leftarrow$ TOP_K_TOPICS(S, I_m, α)
3: select $U^Q_{I_m}$ from U ▷ each $u_i \in U$ has to perform certain number of actions related to Q
4: **for** each $u_i \in U^Q_{I_m}$ **do**
5: compute $\sigma_{(u_i,\psi_{u_i})}$
6: **if** $\sigma_{(u_i,\psi_{u_i})} \geq 0.75$ **then**
7: $\mathcal{C}_\mathcal{H}.add(u_i)$
8: **else if** $(\sigma_{(u_i,\psi_{u_i})} \geq 0.4$ and $< 0.7)$ **then**
9: $\mathcal{C}_\mathcal{M}.add(u_i)$
10: **else if** $(\sigma_{(u_i,\psi_{u_i})} \geq 0.25$ and $< 0.4)$ **then**
11: $\mathcal{C}_\mathcal{L}.add(u_i)$
12: **end if**
13: **end for**
14: **end for**
15: Output the set of topical clusters Φ_Q at each time interval I_m
16: **Procedure** TOP_K_TOPICS(S, I_m, α)
17: $P \leftarrow PriorityQueue(k)$
18: **for** each $T_j \in \mathcal{T}$ **do**
19: compute the total number of mentions $\left|ACTS(*, T_j)\right|$
20: generate user frequency matrix U_{T_j,I_m}
21: compute $\eta_{(T_j,I_m)}$
22: $P.add(\eta_{(T_j,I_m)})$
23: **end for**
24: **return** Top-*k* results from *P*

where $\left|ACTS(*, T_j)\right|$ indicates the total number of activities related to topic T_j and U_{T_j,I_m} represents the number of users who showed their interests on T_j at time interval I_m. The weighting parameter $\alpha \in [0, 1]$ balances the above two factors.

D. Topical Clusters Detection Algorithm

The algorithm, called `Query Algorithm`, identify top-*k* topics from social stream S at each time interval I_m through procedure TOP_K_TOPICS (line 16-24) at first. It computes the trending score $\eta_{(T_j,I_m)}$ for each topic T_j and add that score to a priority queue of size k (line 18-23). Then it returns the top-*k* topics based on their trending scores. Next, the algorithm finds the set of users $U^Q_{I_m}$ from U for a given Q at each time interval I_m and then computes users' interest score $\Omega_{(u_i,\psi_{u_i})}$ (line 3-5). Finally, it groups active users into different clusters based on users' topical interest scores (line 6-11). It outputs Φ_Q at each I_m (line 15). The time complexity of this algorithm is $\mathcal{O}(I_n(q_n + u_n))$. Here I_n, q_n and u_n are the number of time intervals, topics and users in the system respectively.

V. Experimental Evaluation

All experiments are performed on an Intel(R) Core(TM) i5-7220U 2.5 GHz Windows 10 PC with 8 GB RAM. We use a Twitter dataset named SNAP [2] which contains 467 million Twitter posts from 20 million users from June 1, 2009 to December 31, 2009. We randomly choose 4,00,000 users and consider their tweets from June 16, 2009 to June 30, 2009.

A. Experimental Results

In SNAP dataset, we consider users' tweets for query Q consists of top k trending topics. We choose the length of each time window (I_m) as 5 days. At each I_m, we cluster the *active users* (users who having at-least 10 activities related to Q at I_m) 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 scores. The interest score ranges are greater than 0.75, between 0.41 to 0.75 and between 0.25 to 0.4 for $\mathcal{C}_\mathcal{H}$, $\mathcal{C}_\mathcal{M}$ and $\mathcal{C}_\mathcal{L}$, respectively.

We vary the topics in Q for the value of α = 0.50 and use two measures of *entropy* and *cluster topical expertise level* to evaluate the quality of the detected clusters.

$$entropy(\{\mathcal{C}j\}_{j=1}^{r}) = \sum_{j}^{r} \frac{|U(\mathcal{C}_j)|}{|U|} entropy(\mathcal{C}_j) \tag{5}$$

$$entropy(\mathcal{C}_j) = -\sum_{i=1}^{n} p_{ij} log_2 p_{ij} \tag{6}$$

and p_{ij} is the percentage of users in cluster $\mathcal{C}_j$ which are active on the query topic T_i.

$entropy(\{\mathcal{C}_j\}_{j=1}^{r})$ measures the weighted entropy considering all the query topics over all the (r) clusters. Entropy indicates the randomness of topics discussed in clusters. Generally, a good topical cluster should have low entropy value.

Next, we want to measure the semantic cohesion related to Q in each cluster. For this purpose, we identify the main topic of interest of each user u_i according to Equation 7.

$$\lambda_{(u_i,I_m)} = freqmax_Q ACTS(u_i, \psi_{u_i}) \tag{7}$$

Similarly, Equation 8 defines the most frequent topic in a cluster $\mathcal{C}_j$ at time interval I_m.

$$\lambda_{(\mathcal{C}_j,I_m)} = freqmax_Q \lambda_{(u_i,I_m)} \tag{8}$$

Finally, we want to measure the expertise level a cluster is expert (denoted as $\rho_{(\mathcal{C}_j,I_m)}$) for a particular topic T_j at time interval I_m (mentioned in Equation 9).

$$\rho_{(\mathcal{C}_j,I_m)} = \frac{\#\{u_i \in \mathcal{C}_j, \ \ \lambda_{(u_i,I_m)} = \lambda_{(\mathcal{C}_j,I_m)}\}}{|\mathcal{C}_j|} \tag{9}$$

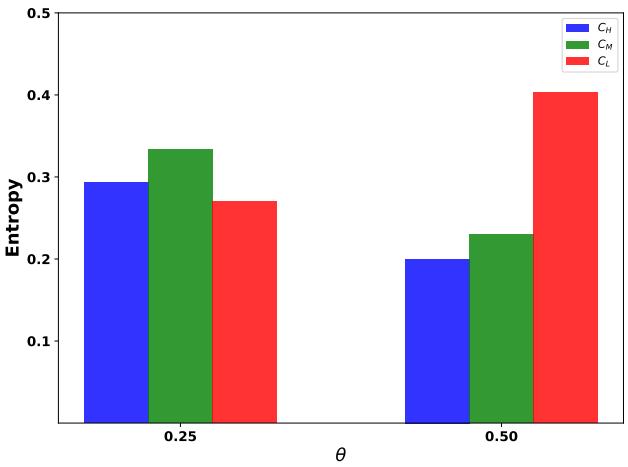

Fig. 3. Entropy at time interval (26/06 - 30/06), $k = 4$, $\alpha = 0.5$

TABLE I
SEMANTIC COHESION ($\rho_{(\mathcal{C}_j, I_m)}$) FOR TOP-$k$ TRENDING TOPICS

Time Window	θ = 0.25	θ = 0.50
I_1 (16 / 06 - 20 / 06)	$\mathcal{C}_\mathcal{H}$ = 0.667 (Business)	$\mathcal{C}_\mathcal{H}$ = 0.647 (Business)
	$\mathcal{C}_\mathcal{M}$ = 0.477 (Business)	$\mathcal{C}_\mathcal{M}$ = 0.563 (Business)
	$\mathcal{C}_\mathcal{L}$ = 0.609 (Business)	$\mathcal{C}_\mathcal{L}$ = 0.486 (Business)
I_2 (21 / 06 - 25 / 06)	$\mathcal{C}_\mathcal{H}$ = 0.750 (Business)	$\mathcal{C}_\mathcal{H}$ = 0.692 (Business)
	$\mathcal{C}_\mathcal{M}$ = 0.406 (Business)	$\mathcal{C}_\mathcal{M}$ = 0.611 (Business)
	$\mathcal{C}_\mathcal{L}$= 0.667 (Business)	$\mathcal{C}_\mathcal{L}$ = 0.633 (Business)
I_3 (26 / 06 - 30 / 06)	$\mathcal{C}_\mathcal{H}$ = 0.50 (News)	$\mathcal{C}_\mathcal{H}$ = 0.538 (News)
	$\mathcal{C}_\mathcal{M}$ = 0.414 (News)	$\mathcal{C}_\mathcal{M}$ = 0.324 (Politics)
	$\mathcal{C}_\mathcal{L}$ = 0.412 (Politics)	$\mathcal{C}_\mathcal{L}$ = 0.333 (Politics)

Fig. 3 shows the entropy values at a particular time interval (26/06 - 30/06) where the trending topics set Q as *Jobs, News, Father's Day and Politics*. For θ values of 0.25 and 0.50 indicate that users have to *active* at-least 1 and 2 topics related to Q respectively. In both cases, we see that entropy values are higher in $\mathcal{C}_\mathcal{M}$ and $\mathcal{C}_\mathcal{L}$ as not all the users show their inclination to all the topics related to Q. On the other hand, as the members in $\mathcal{C}_\mathcal{H}$ have high degree of activeness towards Q, so most of them have to pay attention to all the query topics.

Table I shows the expertise level ($\rho_{(\mathcal{C}_j, I_m)}$) of each cluster for the most frequent topic of that cluster at different time intervals. We find that *Business* is the most frequent topic in all the clusters at time interval I_1 and I_2 for θ values of 0.25 and 0.50, respectively. In all those cases, $\mathcal{C}_\mathcal{H}$ is found as most coherent cluster. In time interval I_3, we see that *News* become most frequent topic (due to the death of *Michael Jackson*) in most cases and cluster $\mathcal{C}_\mathcal{H}$ outperforms other clusters.

In our experiment, we determine trending k topics and observe the changes in different clusters at different time intervals as shown in Table II. The 1st time window we considered for observing the changes following by the change of time windows is considered from 11th June to 15th June. We shift the time window by 5 days and track the changes in second time interval which is from 16th June to 20th June. The top-k ($k = 4$) trending topics are {*Business, Social Media, Father's Day and Politics*} for both first (11/06 - 16/06) and second time interval (16/06 - 20/06). Below are the observed changes:

- 6 members are dropped from $\mathcal{C}_\mathcal{H}$ and 2 members are added to $\mathcal{C}_\mathcal{H}$. Here clusters' members are changed but the cluster size remains same.
- 13 members from $\mathcal{C}_\mathcal{M}$ are dropped and 7 members are added to $\mathcal{C}_\mathcal{M}$.
- 16 members are dropped from $\mathcal{C}_\mathcal{L}$ and 7 members are added to $\mathcal{C}_\mathcal{L}$.

Again we shift the time window by 5 days. The top-k ($k = 4$) trending topics are {*Business, Weather, Father's Day and Politics*} for third time interval (21/06 - 25/06). Below are the observed changes:

- 8 members are dropped from $\mathcal{C}_\mathcal{H}$ and 4 members are added to $\mathcal{C}_\mathcal{H}$.
- 25 members from $\mathcal{C}_\mathcal{M}$ are dropped and 11 members are added to $\mathcal{C}_\mathcal{M}$.
- 27 members are dropped from $\mathcal{C}_\mathcal{L}$ and 22 members are added to $\mathcal{C}_\mathcal{L}$.

Lastly, we make another change for clearer observation of the changes in clusters. We change the time window to observe users' activities from June 25, 2009 to June 30, 2009 and be able to detect {*News, Politics, Father's Day and Jobs*} as trending topic. The observed changes are mentioned below:

- 7 members are dropped from $\mathcal{C}_\mathcal{H}$ and 7 members are added to $\mathcal{C}_\mathcal{H}$. Here clusters' members are changed but the cluster size remains same.
- 12 members from $\mathcal{C}_\mathcal{M}$ are dropped and 28 members are added to $\mathcal{C}_\mathcal{M}$.
- 29 members are dropped from $\mathcal{C}_\mathcal{L}$ and 20 members are added to $\mathcal{C}_\mathcal{L}$.

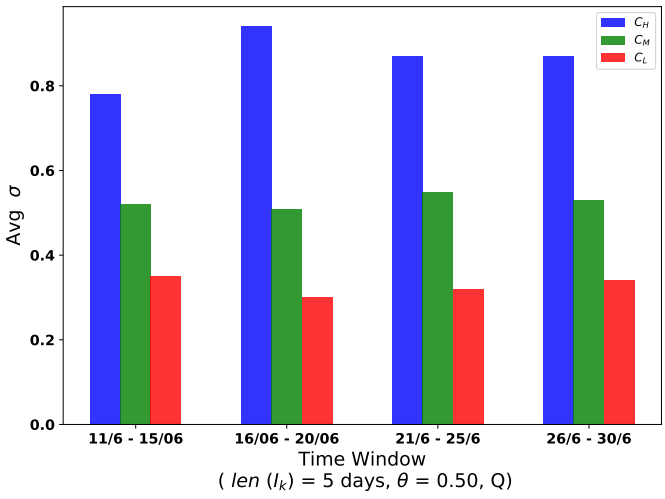

Fig. 4. Average interest score at different time intervals at each topical cluster for top k trending topics where $\theta = 0.5$, $\alpha = 0.5$

Fig. 4 shows the average interest score in each cluster at different time intervals for top k trending topics. We see that the members in high cluster ($\mathcal{C}_\mathcal{H}$) have more average scores in all cases.

TABLE II
TRACKING CLUSTERS AT DIFFERENT I_k FOR TOP-k TRENDING TOPICS IN TWITTER

$I_k \longrightarrow$	(11/6-15/6)	(16/6-20/6)			(21/6-25/6)			(26/6-30/6)		
$\mathcal{C}$	$\|\mathcal{C}\|$	$\|\mathcal{C}\|$	Drop	Add	$\|\mathcal{C}\|$	Drop	Add	$\|\mathcal{C}\|$	Drop	Add
$\mathcal{C}_{\mathcal{H}}$	21	17	6	2	13	8	4	13	7	7
$\mathcal{C}_{\mathcal{M}}$	38	32	13	7	18	25	11	34	12	28
$\mathcal{C}_{\mathcal{L}}$	44	35	16	7	30	27	22	21	29	20

VI. DISCUSSION

In this paper we intend to track user activeness based on the topics which are trending in a particular time interval. We determine clusters and observe the changes that obtain at different time interval.

A. Detection of top-k trending topics in Twitter.

On the basis of user activities on different topics, we discover top-k trending topics successfully. In our 1st time window (11/06 - 15/06) we get (k = *4*) {*Business, Social Media, Father's Day and Politics*} as trending topics. The trending topics remain unchanged in the second time window (16/06 - 20/06). We again change the time window (21/06 - 25/06) and observe that *Weather* topic came to the top-k list as more Twitter users posted tweets related to *Weather* due to excessive heat wave around that period of time. In next time interval (26/06 - 30/06), *News* came to the top list due to the death of popular pop-star Michael Jackson.

B. Tracking changes in cluster for top-k trending topics

For our detected top-k topics at each time interval, we track the following changes in each cluster:

- We observe that by the flow of time, each cluster varies in terms of size and cluster members. Due to different degree of topical interests for different topics at different time intervals, the cluster memberships of some users change at different time windows.
- We determine entropy for each clusters and try to observe the diversity among topics.
- We measure the semantic cohesion related in each cluster. For that we analyse the most active topic for individual user and also for individual cluster.

Detection of trending topic and tracking clusters based on them gives more clear view of people's interest on a particular time interval.

VII. CONCLUSION

The main goal of this work is to cluster analogous Twitter users based on their topical degree of interests over time. It has been observed is that the users' individual activeness vary widely for different attributes. This research outlined an activeness score function for the social users and developed methods to effectively cluster them for top-k trending topics. The effectiveness of the proposed method has been demonstrated over extensive experiments on a real dataset.

REFERENCES

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

[2] J. Leskovec, A. Krevl. SNAP Datasets: Stanford large network dataset collection. (2014)

[3] Han, B., Cook, P., Baldwin, T. Lexical normalization for social media text. In: Journal ACM Transactions on Intelligent Systems and Technology (TIST), Volume 4, Issue 1, pp. 1–27 (2013)

[4] M. Michelson and S. A. Macskassy. Discovering Users' Topics of Interest on Twitter: A First Look.In 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. In Proc. ICDE, pp. 534–545 (2012)

[6] S. Liang, E. Yilmaz, E. Kanoulas. Collaboratively Tracking Interests for User Clustering in Streams of Short Texts. In: 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. In Proc. KDD, pp. 927–936 (2009)