Computer Science and Information Systems 11(1):229–239 DOI: 10.2298/CSIS130220005C

Same Wavelength Group Identification from Online Social Networks: A General Framework

Rafeeque Pandara Chalil[1] and Selvaraju Sendhilkumar[1]

Dept. of Information Science and Technology
Anna University Chennai, India
rafeeqpc@auist.net,thamaraikumar@cs.annauniv.edu

Abstract. Reacting to social issues or events through Online Social Networks has become a social habit. Social scientists have identified several network relationships and dimensions that induce homophily. Sentiments or opinions towards different issues have been observed as a key dimension which characterizes human behaviour. People usually express their sentiments towards various issues. Different persons from different walks of social life may share same opinion towards various issues. When these persons constitute a group, such groups can be conveniently termed same wavelength groups. We propose a novel framework based on sentiments and an algorithm to identify such same wavelength groups from online social networks like twitter. The proposed algorithm generates same wavelength groups in polynomial time for relatively small set of events. The analysis of such groups would be of help in unravelling their response patterns and behavioural features.

Keywords: Same wavelength group, Sentiment analysis, Behavioural analysis, Overlapping community, Homophily.

1. Introduction

In the present information age dominated by communication technologies people resort to innovative ways to express and share their opinions online. A recent statistics[1] show that 76% of twitter users are active tweeters and 23% of facebook users check their account five or more times daily. The phenomenal increase in the volume of user generated content in the form of attitude, opinions, comments etc. in the social media are of immense significance for the analysis of human behaviour.

All Online Social Networks (OSNs) follow the fundamental principle of homophily: similarity breeds connection [19,14]. People in the OSN may be connected to one another with regard to many socio-demographic, behavioural and interpersonal characteristics. Recent studies [12,2,30] show users in the same social circle are more likely to share same opinion. A person's sentiment towards a given issue is determined to a great extent by those of his or her neighbours. For instance, a person's propensity to purchase a commodity is heavily dependent on the kind of opinions likely to emanate from his friends. With this key observation, it is reasonable to state that those who share same sentiments have a strong likelihood of falling into group of similar nature. Such groups would embody persons sharing same opinion on different issues. These persons can be grouped

[1] http://www.socialnomics.net/2012/06/06/

together to form subgroups which can be conveniently termed same wavelength groups. In other words they are the proverbial same feather birds.

Identifying such same wavelength communities online has multifaceted benefits. First, social scientists are enabled to analyse the responses of the group to a socio-political incident or an ethical issue. Second, online recommendation and targeted advertising system can be improved by deep assessment of the groups. Third, responses of the groups can be predicted when a new issue comes up.

Twitter is a micro blogging service in which people share their political, religious, business or personal views in 140 characters not constrained by space and time. Some of the recent works [5,6,31,13] observed that tweet sentiments are strong indicators to predict socio-economic fluctuations. But most of the recent works on twitter sentiments focus either on tweets or the user sentiments on existing groups. We propose a framework[23] and an algorithm to identify same wavelength groups from the public based on the sentiments towards the trending issues or events. The proposed algorithm generates same wavelength groups in polynomial time for a relatively small event-set in a particular time period.The analysis of such groups can unravel the behavioural features and response patterns in a more subtle and effective manner.

The rest of the paper is structured as follows. Section 2 discusses the related works. section 3 discusses the general framework used for the generation of same wave length groups. Section 4 discusses the algorithm and its time complexity. Section 5 examines the experimental results and section 6 concludes with future directions.

2. Related Works

Social scientists have studied extensively the socio-demographic, behavioural and interpersonal characteristics. They used the traditional mode of collecting the data through online, offline and mixed-mode surveys. But recently, the rich data from various OSNs have attracted significant attention from the research community.

Some of the previous work primarily focused on usage statistics and sequences of user activities in OSNs in order to analyse user behaviour. Benevenuto et al. [3] used cickstream data to capture the behaviour of OSN users. They provided a click stream model and observed that silent interactions like profile browsing dominate other visible activities. Guo, Lei et al. [8] analysed users posting behaviour of original content and observed that 20% users contribute 80% total content in the network. Jiang et al. [9] also analyses the latent and visible interactions in OSN. They constructed a latent interaction graph to capture browsing activity among OSN users and observed that latent interactions dominate visible interactions. Lewis et al. [17] created a facebook dataset and they analysed how socio-demographic dimensions like gender, race and ethnicity are correlated with certain network activities. A recent work [20] examined the role of five dimensions (Openness, Conscientiousness, Extroversion, Agreeableness, Neuroticism) of personality on facebook usage and features. They observed that certain personality traits are correlated with facebook usage.

A recent work [11,22] examined how position in the network, activities and user preferences are correlated. They provided a new affinity measure based on distance and conducted studies on email graph and twitter mention graph. They identified the homophily in terms of demography, queries and tweets among the closely connected users. Users in the

OSNs can have multiple affiliations or dimensions. Analysing multiple social dimensions of users exposed to social network environment is known as collective behavioural analysis [27,16]. Behavioural prediction can be made from the learned data model. A recent work [24] used topics, social graph topology and nature of user interactions to discover latent communities in social graphs.

Twitter is micro blogging service to share interesting thoughts at each moment. Most of the recent works on sentiment analysis in twitter [4,5,6,7,10,31] have been done at the tweet level. Some of the recent works [1,26] also considered connected users in the twitter domain to study the behavioural correlation. Tan et al. [26] observed that the probability of sharing the same opinion is high if they are connected. Abbasi et al. [1] have selected an online community which resembles a real world community in terms of race, language, religion etc. They extracted tweets related with Arab Spring to analyse the mood before and after the event.

Communities are not unique and they vary depending on the application of specific needs. Therefore various approaches to identify overlapping communities from social networks have been proposed. Some are based on k-cliques [21,25]. But the above mentioned approaches cannot be applied to a bipartite graph. Biclustering methods are also developed [18,29]. The aforementioned methods extract bipartite clusters and they need not necessarily be bicliques. The same wave length groups are bicliques and in this work the proposed algorithm generates all bicliques from the bipartite graph.

3. General Framework

Same Wavelength Groups are groups formed on the basis of opinions and sentiments of similar hue towards various issues by different individuals. Such same wave length groups vitally connect the individuals in a meaningful and purposeful fraternity. Most of the previous works primarily focused either to analyse sentiments at the tweet level or to study the characteristics of tweeters in a connected environment. But people from different walks of social life may have same opinion on different issues and they need not necessarily be connected. We propose a framework to mine such groups. Fig. 1 shows the general framework for identifying and analysing same wavelength groups.

The tweet extraction phase extracts relevant tweets with respect to the trending issues or events. Public tweets in real-time can be captured with the streaming APIs provided by twitter. Crawling by means of streaming APIs can extract valid and relevant tweets if it is done at the same time when a particular event occurs. Deleting ”news tweets” from corporate tweeters like CNN further refines the extracted collection.

Normalization is fundamental to all text mining tasks. Each extracted tweet may be cryptic and irregular in nature. Moreover tweet may be encoded with a lot of sentiment information like punctuation, emoticons, acronyms etc. So sentiment-aware tokenizing is required to capture emoticons and tweet entities [15]. Normalization phase replaces words having repeating characters (e.g.. loooove), misspelled words and acronyms with proper words or phrases.

Sentiment analyser finds sentiments of users towards each issue or event. Let $U = (u_1, u_2, \ldots, u_m)$ represent the set of users and $E = (e_1, e_2, \ldots, e_n)$ represent the set of events that users respond in a particular time period. Each user may express positive or negative sentiment towards each event. Users sharing the same opinion towards an event

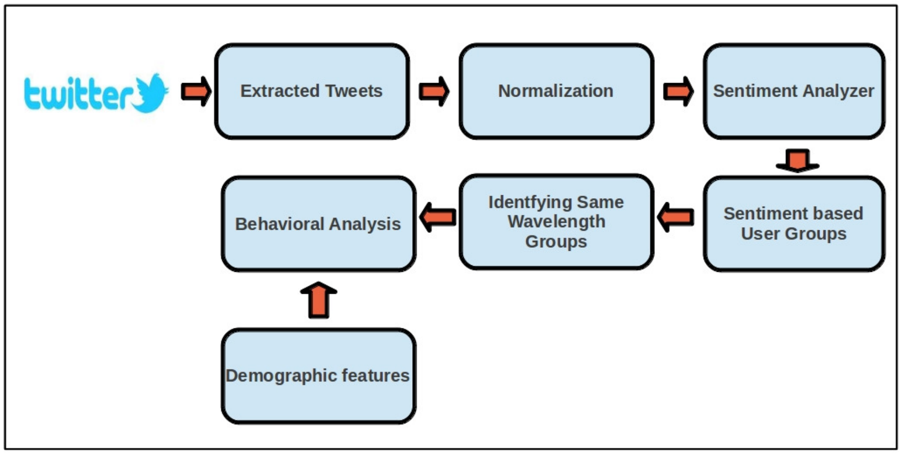

Fig. 1. General Framework

form a $k-clique$ (complete sub graph of size k) where k is the number of users shared the same opinion. For n such events $2n$ such $k-cliques$ will be formed (one for positive and another for negative). Table 1. shows the user-sentiment matrix, say $S_{m\times n}$, where each entry $S(i,j)$ represents the sentiment (positive(P) or negative(N)) towards each event.

Table 1. User-Sentiment matrix $S_{m\times n}$

Users	Event#1	Event#2	...	Event#n
$user_1$	P	P	...	P
$user_2$	P	N	...	N
$user_3$	P	P	...	N
.	...	...	...	...
$user_m$	...	...	...	...

Consider the toy example as shown in Fig. 2. Suppose there are three events (e_1, e_2, e_3) in which nine users $(u_1, u_2, \ldots, u_9)$ express their opinion. The nodes denote users and the edges denote the affiliation with respect to sentiments towards the event. The sets $(u_1, u_2, u_3, u_4, u_5)$, $(u_2, u_3, u_5, u_6, u_7)$ and (u_2, u_3, u_5, u_9) are positive sentiment groups and (u_6, u_7, u_8, u_9), (u_1, u_4, u_5, u_9) and $(u_1, u_4, u_6, u_7, u_8)$ are negative sentiment groups. Each such group form a clique with various sizes. That is an edge $(u_i, u_j) \in clique$ if (u_i, u_j) share the same sentiment towards an event.

Different persons from different walks of social life may share the same opinion towards various issues or events. The dotted line in the toy example shows the common users shared the positive opinion towards three events. if (c_1, c_2, c_3) are three cliques formed from the positive responses towards events (e_1, e_2, e_3) then $(u_2, u_3, u_5, c_1, c_2, c_3)$ form a group which constitutes users and subgroups share the same opinion. This group can be termed same wavelength group.

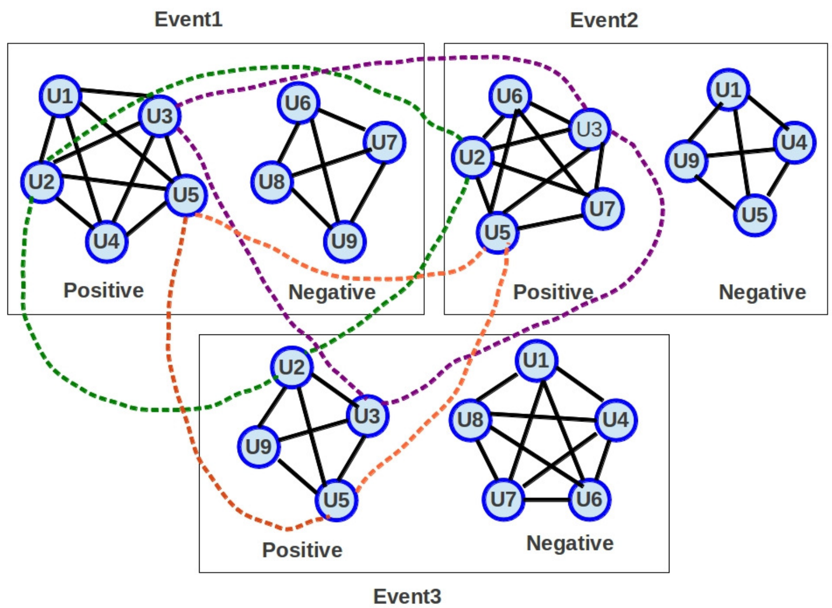

Fig. 2. A toy example. *Each clique represents the sentiments of users towards each event. Dotted line represents the set of users shared the positive opinion towards various events.*

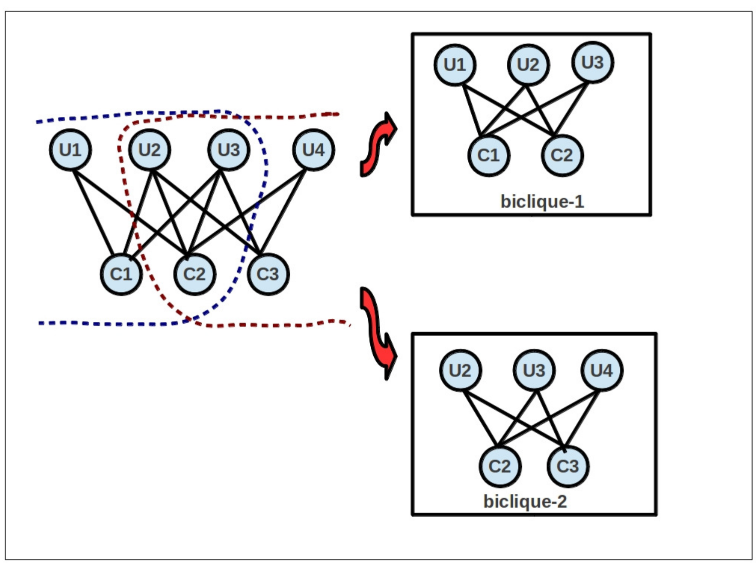

Fig. 3. Identification of same wavelength groups from bipartite graph

Formally, let $U = (u_1, u_2, \ldots, u_m)$ denote the set of m distinct users included in the cliques $c_i (1 \leq i \leq k)$ and $C = (c_1, c_2, \ldots, c_k)$ denote cliques generated based on the opinion towards n events. Now identifying same wavelength groups will reduce to an overlapping community identification problem [21,29] from a bipartite graph $G(U, C, E)$, where U denote set of users and C denote set of groups(cliques) identified by the sentiment analyser phase. For instance, consider a bipartite graph with four users (u_1, u_2, u_3, u_4) and three groups (c_1, c_2, c_3). Fig. 3 depicts how the same wavelength groups can be extracted from a bipartite graph. If the value of k (number of events) is two then three bicliques, $(u_1, u_2, u_3, c_1, c_2)$, $(u_2, u_3, u_4, c_2, c_3)$ and (u_2, u_3, c_1, c_3) can be identified. Fig. 3 shows two bicliques with maximum number of users. In a general bipartite graph with n events, for a particular value of k, $\binom{n}{k}$ such same wavelength can be identified.

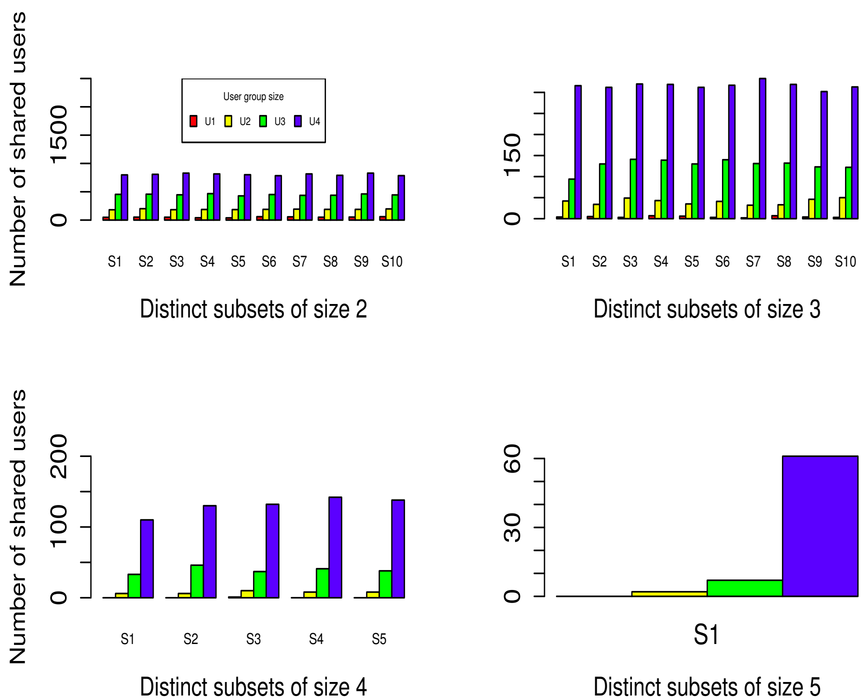

Fig. 4. Number of shared users from distinct subsets of events of varying size

4. Algorithm and Complexity Analysis

The SWG-FIND algorithm first generates $|E|$ cliques of size U_{e_i} where U_{e_i} denotes the number of users having same opinion towards the event e_i. Since the **CREATE-CLIQUE** function depends on creating a complete graph of size U_{e_i} the running time of the same will be $O(U_{e_i}^2)$. For $|E|$ such events (lines 2-4) generates cliques in $|E|\,O(U_{e_i}^2) \approx O(U_{e_i}^2)$ time.

Once the cliques are generated, **CLIQUE-INTERSECT** function finds the intersection of cliques over all subset of events of length greater than one. The total number of such subsets are $2^{|E|} - |E| - 1$. Even though the POWERSET function takes exponential time, fixing $|E|$ as a constant (relatively small value) at a particular instance, the running time may be a constant and perfectly satisfactory.

Algorithm 1 SWG-FIND algorithm

1: **procedure** SWG-FIND(E, U_{e_i}) ▷ $E \leftarrow$ Event-set,$U_{e_i} \leftarrow$ Set of users having same sentiments towards event $e_i \in E$
2:     **for all** event $e_i \in E$ **do**
3:         $C_i \leftarrow$ CREATE-CLIQUE(U_{e_i}, e_i)
4:     **end for**
5:     $S \leftarrow$ POWERSET(E)
6:     **for all** subsets $s_i \in S$ **do**
7:         **for all** $e_i \in s_i$ **do**
8:             **for all** $U_{e_i} \in e_i$ **do**
9:                 $U_{sw_{s_i}} \leftarrow$ CLIQUE-INTERSECT(U_{e_i}, e_i)
10:                 $SWG_i \leftarrow$ CREATE-BICLIQUE($U_{sw_{s_i}}, e_i$)
11:             **end for**
12:         **end for**
13:     **end for**
14:     **return** SWG_i ▷ Same Wavelength Groups
15: **end procedure**

CLIQUE-INTERSECT function returns the set of shared users ($U_{sw_{s_i}}$) having same opinion on different events. The nodes in each clique can be represented as sets. Then the implementation of CLIQUE-INTERSECT function using hash table representation (lines 7-9) runs in $O(n)$ time where n is the total number of users participated in all events in s_i. For $|S|$ such subsets (lines 6-9) finds the intersection in $|S|\,O(n)$ time. Creation of biclique is again creation of complete bipartite graph which runs in $|S|\,O(\left|U_{sw_{s_i}}\right| |e_i|)$ for all subsets of events of length greater than one. The running time of SWG-FIND algorithm is therefore $max(O(U_{e_i}^2), |S|\,O(n) + |S|\,O(\left|U_{sw_{s_i}}\right| |e_i|)$ which is polynomial in time for a fixed and relatively small number of events.

5. Experiment Results and Analysis

We created synthetic data to implement and evaluate the SWG-FIND Algorithm. Synthetic data consist of randomly generated user-ids of sizes 500 (U1), 1000 (U2), 1500 (U3) and 2000 (U4) from a set of 5000 users and event-sets of varying sizes from five to ten. Random user-ids of the above sizes are generated for each event in the event-set. We assume positive sentiments of users participated in each event. POWERSET function generates different subsets of an event-set as described in the algorithm.

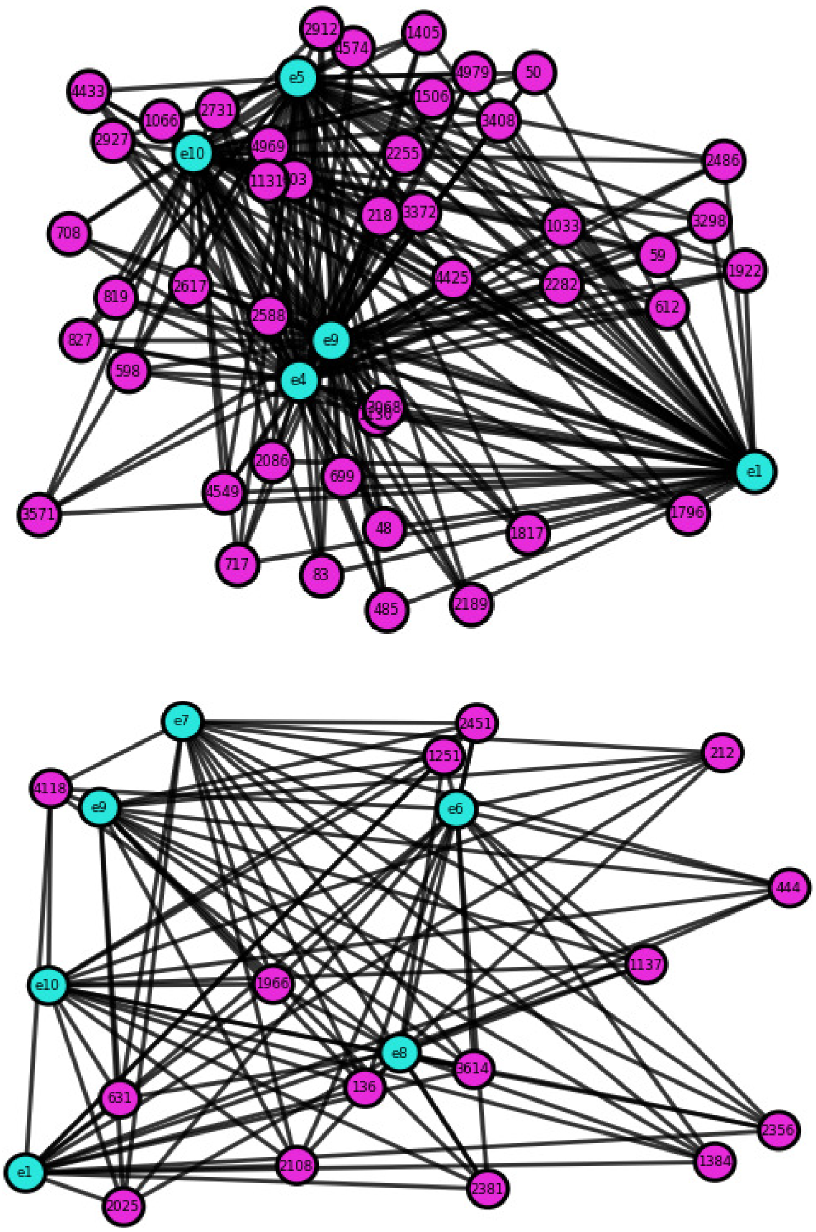

Fig. 5. Two instances out of 1013 bicliques generated for an event-set of size ten

We study how the sentiments of users may vary across subsets of event-set. We considered only subsets of size greater than one since the other case was trivial. Fig. 4 shows the distinct subsets of varying size generated from an event-set of size five. The number of shared users may vary for each subset and also depend on the group (clique) size of each event. There may be cases in which the number of shared users are null since it depends on the number of users in each event as well as the events included in the subset. This may be true in the real data as well since the number of persons reacted may vary across the diverse set of issues or events. Fig. 5 depicts the two instances of bicliques (Same wavelength groups) out of the 1013 bicliques generated for an event-set of size ten. The two distinct kinds of nodes represent event node and user node.

Online social interactions are random and sometimes subtle in nature. Recently Wang, Chunyan and Huberman, Bernardo [28] observed that individual behaviour is less predictable when individuals become members of an explicit group. So Identifying same wavelength group from the public is an extremely more subtle way to analyse the behavioural features. If we closely examine the trending issues or events in which the people react we can very well select the same or diverse nature of event-set of relatively small value in order to analyse the sentiments. Our proposed algorithm generates same wavelength groups of various sizes in polynomial time. This approach can be used for application areas like behavioural modelling, targeted advertising, crowd mood reading, cultural trend monitoring etc.

6. Conclusions and Future Work

Opinions in OSNs have been identified as a strong dimension which induces homophily. In this paper we presented a novel framework for identifying same wavelength groups from online social networks like twitter. The idea is to determine groups of people from the public who share same opinion on various issues or events. This is one subtle way to study the group responses and behavioural patterns. We have mapped the framework to a graph theoretical model and proposed an algorithm which identifies the cliques formed based on the sentiments towards each issue and determines the overlapping bicliques that share the same sentiments towards a set of issues.

This work needs to be explored more using real time twitter data to evaluate the results and computational cost. Moreover human behaviour is dependent on socio-demographic variables like age, sex, education, status etc. Analysing same wavelength groups incorporating socio-demographic features will provide more insights about the evolution of such groups and hence will help to predict future activities. Extending the work including the above features can be a promising future direction.

References

1. Abbasi, M., Chai, S., Liu, H., Sagoo, K.: Real-World behavior analysis through a social media lens. In: 5th International Conference on Social Computing, Behavioral-Cultural Modeling & Prediction (SBP 2012), USA. pp. 18–26 (2012)
2. Adams, J., Faust, K., Lovasi, G.S.: Capturing context: Integrating spatial and social network analyses. Social Networks 34(1), 1–5 (Jan 2012)

3. Benevenuto, F., Rodrigues, T.: Characterizing user behavior in online social networks. In: Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference, NY, USA. pp. 49–62 (2009)
4. Bifet, A., Frank, E.: Sentiment knowledge discovery in twitter streaming data. In: Proceedings of the 13th international conference on Discovery science. pp. 1–15 (2010)
5. Bollen, J., Mao, H., Zeng, X.: Twitter mood predicts the stock market. Journal of Computational Science 2(1), 1–8 (Mar 2011)
6. Bollen, J., Pepe, A.: Modeling Public Mood and Emotion : Twitter Sentiment and Socio-Economic Phenomena. In: Fifth International AAAI Conference on Weblogs and Social Media. pp. 450–453 (2011)
7. Davidov, D., Tsur, O., Rappoport, A.: Enhanced sentiment learning using twitter hashtags and smileys. In: Proceedings of the 23rd International Conference on Computational Linguistics: Posters, PA, USA. pp. 241–249 (2010)
8. Guo, L., Tan, E., Chen, S., Zhang, X., Zhao, Y.E.: Analyzing patterns of user content generation in online social networks. In: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '09. pp. 369–378 (2009)
9. Jiang, J., Wilson, C., Wang, X., Huang, P., Sha, W., Dai, Y., Zhao, B.Y.: Understanding latent interactions in online social networks. In: Proceedings of the 10th annual conference on Internet measurement - IMC '10. pp. 369–382 (2010)
10. Jiang, L., Yu, M., Zhou, M.: Target-dependent twitter sentiment classification. In: 49th Annual Meeting of the Association for Computational Linguistics, Oregon, June. pp. 151–160 (2011)
11. Jung, J.J.: Boosting social collaborations based on contextual synchronization: An empirical study. Expert Systems with Applications 38(5), 4809–4815 (2011)
12. Jung, J.J.: Service chain-based business alliance formation in service-oriented architecture. Expert Systems with Applications 38(3), 2206–2211 (2011)
13. Jung, J.J.: Attribute selection-based recommendation framework for short-head user group: An empirical study by movielens and imdb. Expert Systems with Applications 39(4), 4049–4054 (2012)
14. Jung, J.J.: Collaborative browsing system based on semantic mashup with open apis. Expert Systems with Applications 39(8), 6897–6902 (2012)
15. Jung, J.J.: Online named entity recognition method for microtexts in social networking services: A case study of Twitter. Expert Systems with Applications 39(9), 8066–8070 (2012)
16. Jung, J.J.: Cross-lingual Query Expansion in Multilingual Folksonomies: a Case Study on Flickr. Knowledge-Based Systems 42, 60–67 (2012)
17. Lewis, K., Kaufman, J., Gonzalez, M., Wimmer, A., Christakis, N.: Tastes, ties, and time: A new social network dataset using Facebook.com. Social Networks 30(4), 330–342 (Oct 2008)
18. Madeira, S.C., Oliveira, A.L.: Biclustering algorithms for biological data analysis: A survey. IEEE/ACM Trans. Comput. Biol. Bioinformatics 1(1), 24 – 45 (2004)
19. Mcpherson, M., Smith-lovin, L., Cook, J.M.: BIRDS OF A F EATHER : Homophily in Social Networks. Annual Review of Sociology 27(1), 415–444 (2001)
20. Moore, K., McElroy, J.C.: The influence of personality on Facebook usage, wall postings, and regret. Computers in Human Behavior 28(1), 267–274 (Jan 2012)
21. Palla, G., Derényi, I., Farkas, I., Vicsek, T.: Uncovering the overlapping community structure of complex networks in nature and society. Nature 435(7043), 814–818 (Jun 2005)
22. Panigrahy, R., Najork, M., Xie, Y.: How user behavior is related to social affinity. In: Proceedings of the fifth ACM international Conference on WSDM'12, Washington. pp. 713–722 (2012)
23. Rafeeque, P.C., Sendhilkumar, S.: Identifying same wavelength groups from twitter: A sentiment based approach. In: ACIIDS 2013, Kuala Lumpur, Malaysia,Intelligent Information and Database Systems, Lecture Notes in Computer Science, vol. 7803, pp. 70–77. Springer Berlin Heidelberg (2013)

24. Sachan, M., Contractor, D., Faruquie, T.a., Subramaniam, L.V.: Using content and interactions for discovering communities in social networks. In: Proceedings of the 21st international conference on World Wide Web - WWW '12. pp. 331–340 (2012)
25. Shen, H., Cheng, X., Cai, K., Hu, M.B.: Detect overlapping and hierarchical community structure in networks. Physica A: Statistical Mechanics and its Applications 388(8), 1706 – 1712 (2009)
26. Tan, C., Lee, L., Tang, J., Jiang, L., Zhou, M., Li, P.: User-level sentiment analysis incorporating social networks. In: ACM International Conference on Knowledge and Data Engineering (KDD'11), California, USA. pp. 1397–1405 (2011)
27. Tang, L., Liu, H.: Toward Collective Behavior Prediction via Social Dimension Extraction. IEEE Intelligent Systems 25(6), 19–25 (2010)
28. Wang, C., Huberman, B.a.: How Random are Online Social Interactions? Scientific reports 2, 633–638 (Jan 2012)
29. Wang, X., Tang, L., Gao, H., Liu, H.: Discovering Overlapping Groups in Social Media. In: 2010 IEEE International Conference on Data Mining. pp. 569–578 (Dec 2010)
30. Yang, X., Steck, H., Liu, Y.: Circle-based recommendation in online social networks. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 1267–1275 (2012)
31. O'Connor, B.: From tweets to polls: Linking text sentiment to public opinion time series. In: Proceedings of the International AAAI Conference on Weblogs and Social Media, Washington, DC. pp. 122–129 (2010)

Rafeeque Pandara Chalil is a corresponding author of this paper. He is an Associate professor in Computer Science and Engineering at Govt. College of Engineering Kannur, Kerala, India since 2009 and currently working toward the PhD degree in Information and Communication at Anna University Chennai, India. He received B.Tech in Computer Science and Engineering from Calicut university, Kerala , India in 1996 and the masters degree from National Institute of Technology Calicut, India in 2007. His research interests are Social computing, Text/Web mining. Specifically he is interested in sentiment and behavioural analysis in online social networks.

Selvaraju Sendhilkumar is an Assistant Professor (Sr. Gr.) in Anna University, Tamilnadu, India since January 2004.He received the B.Tech. in Electrical & ElectronicsEngineering from Pondicherry University in 2000. He received M.Tech. in Information Technology and Ph.D. degree in Computer Science & Engineering fromAnna Universityin 2002 and 2009, respectively. His research topics are Web Mining, application of Web mining in social networks andcitationanalysis, Big DataAnalytics.

Received: February 20, 2013; Accepted: September 11, 2013.