Understanding User-Community Engagement by Multi-faceted Features: A Case Study on Twitter

Hemant Purohit
Kno.e.sis, Dept. of Computer Science and Engineering
Wright State University
hemant@knoesis.org

Yiye Ruan
Dept. of Computer Science and Engineering
Ohio State University
ruan@cse.ohio-state.edu

Amruta Joshi
Dept. of Computer Science and Engineering
Ohio State University
joshiam@cse.ohio-state.edu

Srinivasan Parthasarathy
Dept. of Computer Science and Engineering
Ohio State University
srini@cse.ohio-state.edu

Amit Sheth
Kno.e.sis, Dept. of Computer Science and Engineering
Wright State University
amit@knoesis.org

ABSTRACT

The widespread use of social networking websites in recent years has suggested a need for effective methods to understand the new forms of user engagement, the factors impacting them, and the fundamental reasons for such engagements. We perform exploratory analysis on Twitter[1] to understand the dynamics of user engagement by studying what attracts a user to participate in discussions on a topic. We identify various factors which might affect user engagement, ranging from content properties, network topology to user characteristics on the social network, and use them to predict user joining behavior. As opposed to traditional ways of studying them separately, these factors are organized in our framework, *People-Content-Network Analysis (PCNA)*, mainly designed to enable understanding of human social dynamics on the web. We perform experiments on various Twitter user communities formed around topics from diverse domains, with varied social significance, duration and spread. Our findings suggest that capabilities of content, user and network features vary greatly, motivating the incorporation of all the factors in user engagement analysis, and hence, a strong need can be felt to study dynamics of user engagement by using the PCNA framework. Our study also reveals certain correlation between types of event for discussion topics and impact of user engagement factors.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data Mining

General Terms

Human Factors, Languages, Experimentation

Keywords

Social Networks, Community Formation, User Engagement, Twitter, Content Analysis, Network Analysis, People-Content-Network Analysis (PCNA)

1. INTRODUCTION

Social media has revolutionized the way of user interaction with information. Social network users are not only creators and recipients of the information, but also critical relays to propagate information. The powerful ability of sharing has played an important role in events with varied social significance, audience, and duration, such as political movements (e.g. Jasmine Revolution), brand management and marketing[2], and emergency management (e.g., Haiti, Japan earthquake).

This shift in the paradigm of information creation and consumption has presented to all social entities a challenge for better understanding the type and level of user engagement. Here authors consider user engagement definition as user joining a community surrounding topic discussion on social network by writing or sharing messages about that topic. The knowledge of user participation behavior has a number of compelling applications. A first example is movie studio's strategy making on spreading the message of a movie's release in social media. If they can identify prominent factors affecting user engagement, those factors can be emphasized accordingly to maximize the word-of-mouth effect. In another use case, during an event of crisis, emergency teams are looking forward to help the victims. User engagement analysis could help us understand how effectively the community surrounding this event can grow to reach potential donors and people in need of resources (food, water, first aids etc.), also what are the best possible ways to communicate between these resource providers and people in need for resources etc.

[1] `http://www.twitter.com`

[2] `http://www.chromaticsites.com/blog/impressive-twitter-customer-service-brand-management-cases`

The study of user engagement is by no means simple, as there is a three-dimensional dynamic at play: topic of interest (*content*), participants (*people*) who engages in discussion about the topic and community (*network*) formed around the topic. Researchers have been addressing this problem on different facets, like, design and theory of online communities[15], social network data analytics[8, 4], information propagation[7, 16, 21, 18], community detection[1, 10, 22, 9], and link prediction[2, 17, 3].

In this paper, we focus on one key question: given a discussion topic on social media, what motivates a user to engage in the discussion for his/her first interaction? Here a topic is formalized as a real-world event, discussions are thus surrounding this event, and all participants compose a community (which will be formally defined as *event-oriented community* in section 3.2). For example, during Japan Earthquake 2011, an event of natural disaster, people tweeting about *Japan Earthquake* would be considered to be part of the *Japan Earthquake* topic discussion community. The task of finding factors which drive user to engage in topic discussion, therefore, can be considered as identifying factors that influence user to join the corresponding community. We use Twitter as a social information source and manage to build a prediction model for user engagement in topic discussion about events. Compared with previous related works which resort to small subsets of features, and isolated study of factors with different characteristics, we investigate a range of features in three categories (content, author, and community/ network) and build an organized framework, namely *People-Content-Network Analysis (PCNA)*, for studying the factors responsible for user engagement on social media.

Our experiments on Twitter event-oriented communities demonstrate that capabilities of content, author and network features vary greatly for impacting user engagement, motivating the incorporation of all the factors, and hence, a need to study dynamics of user engagement by using the PCNA framework. Our findings through prediction model performances suggest that content features contribute most for influencing user to engage in topic discussion, followed by people and network characteristics for most of the discussions of topics. Moreover, we find correlations between event types and features, which can help understand user engagement in better scientific ways.

The paper is organized as, review of past works in the related topics in section 2, a formal definition of the problem of interest in section 3 and description of all features considered and methods used in section 4, followed by experimental results and analysis in section 5. Finally we conclude the findings from present work and list future work directions in section 6.

2. RELATED WORKS

Researchers have been studying social networks to understand the dynamics of user engagement in various forms, such as community formation, community detection, information propagation, link prediction, etc.

On the topic of community formation and community detection, Backstrom et al. proposed a model for network membership, growth and evolution [2] by analyzing DBLP and LiveJournal social networks. They found that the propensity of individuals to join communities, and of communities to grow rapidly, depended in subtle ways on the underlying network structure. Shi et al. studied the patterns of user participation behavior and feature factors that influence such behavior on four web forums [17]. Their results of BiMRF modeling with two-star configurations suggested that user similarity defined by frequency of communication or number of common friends was inadequate to predict grouping behavior, but adding node-level features could improve the fit of the model. Leicht and Newman introduced a solution to find communities in directed networks [9]. They showed how modularity maximization could be generalized in a principled fashion to incorporate directionality of the graph. Leskovec et al. studied clustering problem on a wide range of real-world large networks [10] and concluded that ideal size for most community-like clusters was around 100 nodes.

For previous works on information propagation, Lin et al. suggested a model for tracking popular events in online social network [12] by focusing on the interplay between textual content and social networks. Specifically, they defined a Gibbs Random Field to model the influence of historical status of actors in the network and the dependency relationships among them; thereafter a topic model generated the words in text content of the event, regularized by the Gibbs Random Field. Suh et al. presented extensive analysis on retweeting behavior on Twitter while identifying important content features responsible for attracting new users in the diffusion chain [18]. Nagarajan et al. studied viral content on Twitter and finds out a clear relationship between sparse/dense retweet patterns and the content and type of a tweet itself [13], suggesting the need to study content properties in link-based diffusion models. Romero et al. proposed an algorithm that determined the influence and passivity of users based on their information forwarding activity [16]. It suggested that both measures were important to understand user engagement, as for individuals to become influential they must not only obtain attention and thus be popular, but also overcome user passivity.

Liben-Nowell and Kleinberg surveyed various unsupervised methods on the link prediction problem [11] and conducted extensive experiments on co-authorship networks. More recently, Backstrom and Leskovec introduced Supervised Random Walk [3] to solve link prediction problem, which combined information from the network structure with node and edge level attributes to learn a function that assigns strengths to edges in the network such that a random walker will visit the nodes to which new links will be created in the future.

The discussion above reveals one problem among the approaches taken so far: network, content and user features have been studied in isolated ways. On the other hand, our methodology combines the network characteristics, people/user characteristics at node level, in addition to content level features which forms the basis for topic community. Those three groups of features are therefore leveraged within the PCNA framework to understand user engagement, comprising advantages over previous methods based on fewer feature dimensions.

3. PROBLEM STATEMENT

3.1 Terminology Definition

As described in introduction, user engagement in a topic discussion can be understood in terms of user participation in community formed around topic of discussion; we define

some terminologies used in the context here and then give our problem statement:

- **Event-Oriented Community:** We define an event-oriented community as an implicit group of social network users who have joined discussion on topic about an event, or more precisely who have posted messages about the topic. In different online social networks, posts may appear in different forms (e.g. *status, share* and *comment* for Facebook or *tweet* and *retweet* for Twitter). A social network user is considered to become engaged in the topic discussion and hence, a member of the event-oriented community if it writes or forwards the event-related post; e.g., all the twitterers who are posting about Emmy Awards and thus joining topic of discussion during Aug 10 to Sept 20, 2010 are regarded as members of Emmy Awards community.
- **Slice and Snapshot:** A slice refers to the collection of event-related messages (tweets) or in other words, messages relevant to topic of discussion, posted during a fixed-length period of time (e.g. 24 hours). A snapshot refers to state of the network at a certain point of time at which user profile and connection information are stored. In the current context, we take the snapshots for the network of users who are members of the community formed around topic of discussion.
  Depending on the inherent characteristics of event, we set two different slice lengths (one day and eight hours, respectively) in order to capture the dynamics of community more promptly, since some event-oriented topics of discussion draw a quick attention of users, which in turn engages huge number of users. More detailed discussion is available in section 4.2.
- **Temporal Weight of Information:** While the total size of community surrounding topic of discussion keeps increasing as it evolves, the *freshness* of it should also be taken into account when we study users' behavior. Most online social networks' layout designs show the latest information first, and users have to scroll down to see earlier news feed. Therefore, it is natural to assume that later the information is generated, the higher possibility it get consumed [20] and the higher weight it carries on influencing user decision about engaging in topic discussion.
  To leverage this observation, we set a hard margin of 3 slice units and only consider information tweets within this time window as we believe they are most likely to be viewed. Users who wrote or shared event-related messages during this period are called *active users*, and we would like to focus on how they joined the event-oriented communities, forming the *active community*, and how their followers (i.e. audience) will react in accordance. For each active user and the content he/she generated, a temporal weight is leveraged based on the time that has elapsed since its creation.

Figure 1 illustrates the notions of slice, snapshot and active community, to provide the readers a clearer conception.

3.2 Problem Definition

Using the terminologies introduced so far, the problem of finding factors impacting user engagement can be defined as

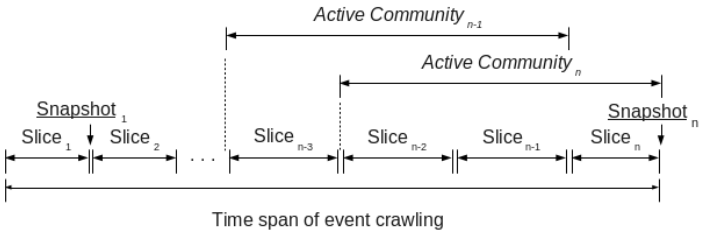

Figure 1: Illustration of slice, snapshot and active community

user engagement prediction problem for joining a topic of discussion:

Def. 1. User Engagement Prediction Problem*: Given 1) an event-oriented community* $\boldsymbol{C}$ *formed around a topic of discussion; 2) a Twitter user* $\boldsymbol{U} \notin \boldsymbol{C}$*, predict whether* $\boldsymbol{U}$ *will be engaged in* $\boldsymbol{C}$ *(by composing a new tweet or retweeting an existing tweet which contains keywords or hashtags related to* $\boldsymbol{C}$*'s underlying event) in a future slice. If so,* $\boldsymbol{U}$ *is said to be a positive record. Otherwise, it is a negative record.*

4. METHODS

This section introduces methodologies involved in building prediction model. Particularly, a detailed discussion about the groups of features used to build the model in order to solve the user engagement prediction problem as defined in section 3.2 is provided. We also describe organization of various features in our *People-Content-Network Analysis (PCNA)* framework.

4.1 Twitter as a Data Source

Launched in 2006, Twitter has been well-known both as a micro-blogging service provider and a social network platform. A message posted by the user is called a *tweet*, which typically contains plain texts, *hashtags* (For example; #nsn3d, #MusicMonday) that indicate explicit topic categorization and hyperlinks to other multi-media content that promote spread of information from all over the Web. The length of each tweet is limited to 140 characters.

Users of Twitter have directed *follower* connections with other users of the site that allows them to keep track or *follow* those other users. Members can post tweets, respond to a tweet which is called *reply* or forward a tweet to all followers which is called *retweet*. Replies to any tweet are directed to a user (not the conversation thread) by placing a *@user* reference in the tweet while retweets are means of participating in a diffuse conversation. The *@user* reference can also be used to refer to a particular user in the tweet content, which is called a *mention* of that user.

Tweets are generally available as feeds from follower networks and also via a searchable interface. Apart from the 140-character text itself, timestamp and location information are all publicly retrievable unless privacy control is turned on by the user. We store most of this data to construct features.

To investigate the users' behavior after perceiving activities of the community, it is not feasible to randomly pick users from millions of Twitter accounts as it is not clear if they are aware of the event at all. Instead, all active users

and the users who follow at least one active user at that snapshot are considered. There are two indications here. Firstly, most users are guaranteed to have access to the event information from the topic related tweets posted or retweeted by their online friends. Thus sophisticated social network features can be used to analyze information propagation in networks. Secondly, a user may not be active for several snapshots before joining the community, resulting in many *negative* records and one *positive* record. The collection of all records and the edges joining them form an *active network*.

4.2 Community Categorization by Event Characteristics

Popular events on social networks belong to diverse domains and differ in characteristics and behavior. Some events like FIFA World Cup drive attention of global populace, while Health Care debate events are of national interest and few other events similar to Iowa State Fair are attractive to a relatively small region. Another categorization is based on the event occurrence and duration. FIFA World Cup event is scheduled long time in advance while events such as Earthquake in Haiti has sudden occurrence. Apparently the characteristics of an event-oriented community largely depends on the triggering event, and it is intriguing to explore the relation of user participation behavior with communities' nature. We expect a variety of community gathering around events and one characteristic from each of the following categorizations is assigned to each event-oriented community (see section 5.1 for details):

- ***Global vs. Local:*** Depending upon the interest level an event can be *global* (such as Emmy Awards) or *local* (such as Iowa State Fair) communities. *Local* communities can further be distinguished by national interest (For example, fans of NFL championship in US) and regional interest (For example, Ohio State Fair in Ohio), though it is not explicitly specified in the present work.
- ***Compact vs. Loose:*** Events may interest varying audiences within which the level of existing interaction among users changes significantly. For example, two fans mentioning the release of a new movie may not have talked to each other previously at all, thus the community formed around this topic is very *loose*. Meanwhile, interested authors for a technical conference topic like LinuxCon are highly likely to have interacted with each other before, and therefore belong to a relatively *compact* community.
- ***Deterministic vs. Unexpected:*** A few events are known to us beforehand while others have sudden occurrence. Therefore, the corresponding communities are *deterministic* and *unexpected*, respectively.
- ***Transient vs. Lasting:*** Different events create different level of buzz in the community and so the community might be either *transient* or *lasting*. As an example of *transient* community, there was hardly anyone talking about the hostage incident in Discovery Channel Building in Seattle three days after it since the crisis was resolved within hours. Meanwhile, discussion of the movie Avatar lasted for months. For the *transient* communities, a unit length of eight hours for time slice is used to capture fast-changing trends better. For the *lasting* communities, we use one day as the unit length for time slice.

4.3 Feature Categorization

Previous works have employed a wide range of features which generally fall into three categories: community, author and content. Those works, however, seldom incorporated multiple groups of knowledge into a single model. We organize these features in our framework, *People-Content-Network Analysis (PCNA)*, and investigate which ones contribute most to the predictability.

4.3.1 Community Features

Community features involve several measurements of the event-oriented community including the size of the active community, the total number of active users that $\boldsymbol{U}$ is following, the size of the weakly connected component (WCC) in the active network that $\boldsymbol{U}$ belongs to, and the ratio of this WCC's size to the active network's size.

4.3.2 Author Features

Author features involve statistics about the active users that $\boldsymbol{U}$ is following, as they are the main source of $\boldsymbol{U}$'s awareness and knowledge of the topic. We would like to discover if those users' social network states have any influence on $\boldsymbol{U}$'s participation behavior. We consider the counts of followers, followees as the features since they implicitly reflect authors' influences.

The influence and passivity scores proposed by Romero et al. [16] can also be meaningful author features. However, the original Influence-Passivity algorithm requires the appearance of hyperlinks in each tweet, which may not fit well in the current scenario. As an alternative, a composite score called *klout*[3] takes most of those measures into account and is publicly available. Therefore, each author's klout score is included in author features.

Moreover, not all users are equally active. As described in section 3.1, temporal weight is applied to author features, so the values are weighted w.r.t. the elapsed time since his last activity in the community (i.e. writing or sharing a tweet related to the topic).

4.3.3 Content Features

Content analysis in the context of social network is more than pure language analysis as information is conveyed in a variety of formats. As a result, number of occurrences of platform-specific features for Twitter (retweet, mention, hashtag) as well as relevant keywords are kept track of.

Hyperlinks in tweets also play an important role in the process of information diffusion, as the content of external pages that is referred to can build better context for the topic of discussion and may motivate $\boldsymbol{U}$. In our practice, each tweet can either have a relevant link, an irrelevant link or no link. To determine whether a link is relevant, we rely on searching for event keywords in the webpage that the link points to. If there is a match, the link is considered relevant; otherwise it is irrelevant. The count of hyperlinks in each tweet is therefore adjusted to 1, -1 and 0 for the three cases, respectively.

We also compute the extent of subjectivity of those tweets as part of the content features. The reason is that we can study if there is any preference of objective, fact-sharing

[3] `http://klout.com`

messages to subjective, emotional messages in terms of information propagation and thus attracting user to the community. As measuring subjectivity is a non-trivial task in the study of natural language processing [14], a simple heuristic is designed, focusing on two groups of explicit features. The key components used towards the score calculation are the subjectivity of (word, part-of-speech tag) pairs and that of emoticons found in the tweet. For the former, we start by feeding tweets into a part-of-speech tagger [19], keep all content words (noun, verb, adjective or adverb) and then classify those word-tag pairs using a pre-compiled subjectivity lexicon[4]. Entries in the lexicon are labeled as either strongly subjective or weakly subjective, and we assign 2 points to each strongly subjective pair, 1 point to each weakly subjective pair and 0 point otherwise. For the latter component, we compiled a lexicon[5] which holds more than 130 commonly-used emoticons. The scoring scheme for emoticon is the same as that for word-tag pair. The final subjectivity score for a tweet m is computed as the average of those segments' scores:

$$S_{score}(m) = \frac{\sum_{(w,t)\in WT(m)} subj_{pair}(w,t) + \sum_{e\in EMOT(m)} subj_{emot}(e)}{|WT(m)| + |EMOT(m)|}$$

where $WT(m)$ is the list of word-tag pairs in m, and $EMOT(m)$ is the list of emoticons in m.

Content analysis is further enriched by linguistic cues in text, which are extracted from analysis through Linguistic Inquiry and Word Count (LIWC)[6] dictionary. LIWC provide statistics of words grouped by grammatical (e.g. preposition) or semantic (e.g. words that describe an *occupation*) components. We apply Principle Component Analysis (PCA)[7] to find out top 3 features in the LIWC analysis results, which are included as content features.

Moreover, as described in section 3.1, temporal weight is applied to content features. Here content features are computed for content posted by active friends of $\boldsymbol{U}$.

4.4 Model Fitting

As there are two possible outcomes of user participation behavior and all the aforementioned features take real values, we treat the User Engagement in Topic Discussion Problem as a binary classification problem operated on feature vectors of the following format.

- label: fact of whether the user joining the community or not. The value for is binary variable can be either positive or negative, and it serves as the class label.
- Community Features:
  - *wccSize*: size of the WCC which the user belongs to in the active network.
  - *wccPercent*: ratio of the WCC's size to that of the whole active network.
  - *connectivity*: number of active friends (i.e. followees) in the active community.
  - *communitySize*: size of the active community.
- Author Features:
  - *logFollower*: logarithm of the weighted geometric mean of active friends' counts of followers.
  - *logFollowee*: logarithm of the weighted geometric mean of active friends' counts of followees.
  - *klout*: weighted means of active friends' klout scores.
- Content Features:
  - *url, retweet, mention, hashtag, keyword*: weighted means of the counts of relevancy-adjusted url, retweet, mention, hashtag, keyword in all active friends' tweets.
  - *sentiment subjectivity*: weighted mean of sentiment subjectivity score.
  - *pca1, pca2, pca3*: weighted means of the top 3 PCA features on LIWC results applied to all active friends' tweets.

The temporal weight vector is set as $(1, 0.8, 0.6)$. That is, assuming the current slice of consideration is slice k, the temporal weight for each tweet is 1, 0.8, or 0.6 if it was written in slice k, $k-1$ or $k-2$, respectively. Any tweets written earlier than two slices ago are no longer considered active.

Algorithm 1 describes the pseudo-code for generating all records for the classification problem.

5. EXPERIMENTS

5.1 Data Collection

Tweet stream for topics was crawled with Twitter's Search API[8] using an initial seed of manually compiled keywords and hashtags relevant to the event. For a keyword k, we crawl all tweets that mention k, K, #k and #K. The seed list of keywords and hashtags is kept up-to-date by first automatically collecting other hashtags and keywords that frequently appear in the crawled tweets and then manually selecting highly unambiguous hashtags and keywords from this list. We avoid the query drift problem by placing a human in the loop to ensure that ambiguous keywords are not crawled outside of context but only in combination with a contextually relevant keyword.

Data crawl was performed at fixed time intervals, here every 30 seconds. For every issued query, the Twitter search API responds with 1500 tweets. Crawling at regular and frequent intervals allows us to make an assumption that the data collected is a close approximation of the actual population of the tweets generated for the event in that time period. We also crawl the social graph (i.e. follower list) of these tweet posters, who are part of this event-oriented community at specific timestamps of the day. Duration for the time gap between subsequent snapshots of the network for different communities depend on the type of event. We also collect tweet posters' profile information like location, followers and followees counts, description about the tweet poster, etc. For those users who activated privacy setting, no information was crawled, and their tweets are discarded from the slice.

A total of 14 events are considered, and information of these communities are crawled. They were popular topics

[4] http://www.cs.pitt.edu/mpqa

[5] http://www.cs.umbc.edu/courses/331/spring10/2/hw/hw7/hw7/data/sentislang.txt
http://en.wikipedia.org/wiki/List_of_emoticons

[6] http://www.liwc.net

[7] Modified from http://www.neuroshare.org

[8] http://apiwiki.twitter.com/w/page/22554756/Twitter-Search-API-Method:-search

Events	#tweets	#unique users	%relevant url	%mention	%RT	%emoticons	average. subj. score	average active community size	average connectivity
ClevelandShowPremiere	1494	1221	19.26	25.97	11.85	6.16	0.28	686.23	1.28
DiscoveryBuildingCrisis	3303	2580	48.97	12.14	35.06	5.60	0.19	1497.87	2.67
EmmyAwards	5027	3453	65.12	11.06	17.57	6.47	0.18	1126.39	3.10
GoogleInstantSearch	4058	3429	63.05	9.78	23.48	4.09	0.14	1611.79	3.32
HeismanTrophy	5631	4261	32.23	9.06	33.17	2.26	0.16	2487.05	2.61
IowaStateFair	2470	1106	33.59	36.92	21.05	8.54	0.20	349.72	4.83
JewishNewYear	7676	6251	17.18	19.72	25.68	9.64	0.23	3097.96	2.51
LindsayLohanHearing	5547	3660	55.39	6.99	27.08	3.19	0.13	1210.49	1.95
LinuxCon	1294	418	36.14	18.86	33.69	4.71	0.17	226.85	3.11
LondonTubeStrike	1186	530	56.70	15.00	18.47	10.96	0.15	161.6	1.35
RichCroninDeath	476	379	25.16	30.46	25.42	18.70	0.24	215.06	1.16
ScottPilgrimRelease	21435	14286	31.30	13.21	21.63	8.84	0.17	3979.91	2.79
SESSanFrancisco	1383	462	85.62	5.28	10.48	4.99	0.09	157.89	2.59
StuxnetWorm	2845	1855	70.91	8.19	21.83	6.40	0.17	1458.85	3.29

Table 1: Statistical summarization for data sets

Algorithm 1: Generating all classification records

```
timeWgt ← (1.0, 0.8, 0.6)  // Temporal weight
winLen ← 3          // Active window length
def selfUnion(Set P, Set P′):// Auxiliary function
begin
  P = P ∪ P′
end
def label(User U, Slice S):  // Auxiliary function
begin
  if U ∈ activeCommunity[S.id] then return "pos" else
  return "neg"
end
def makeAllRecord(Dataset D):  // Main function
begin
  foreach Slice S ∈ D do
    foreach Author A ∈ S do
      selfUnion(activeCommunity[S.id], {A})
    end
    foreach int I ← 1 to min(winLen − 1,S.id) do
      selfUnion(activeCommunity[S.id − I],
        activeCommunity[S.id])
    end
  end
  foreach Slice S ∈ D do
    foreach Author A ∈ S do
      foreach int I ← 0 to
      min(winLen − 1,D.size − S.id − 1) do
        selfUnion(activeNetwork[S.id + I],{A} ∪
        A.followers)
        foreach User F ∈ A.followers do
          selfUnion(F.activeFriends[S.id + I],{A})
          foreach Tweet T ∈ A.tweets[S.id ] do
            selfUnion(F.partialRecords[S.id + I],
              {timeWgt[I] ×
            (A.features[S.id], T.features)})
          end
        end
      end
    end
    foreach User U ∈ activeNetwork[S.id] do
      print ((U.id, label(U,S),
        activeNetwork[S.id].wccSize[U],
        activeCommunity[S.id].size,
        U.activeFriends[S.id].size,
        avg(U.partialRecords[S.id]))
    end
  end
end
```

(i.e. buzz words) at the period of crawling[9], showing steady growth in number of related tweets in the real-time search result. Furthermore, they are all social events with impacts beyond the online world. For most of predefined events the crawl was started in advance and extended after the event duration. The following list introduces each event and its categorization as defined in section 4.2. Due to the space constraint, only a summarization is provided.

- **ClevelandShowPremiere**: Second Season premiere of animated TV series *Cleveland Show*. September 26. Global, loose, deterministic, transient.
- **DiscoveryBuildingCrisis**: Hostage crisis at the headquarters of Discovery Channel, Maryland. September 1. Local, loose, unexpected, transient.
- **EmmyAwards**: 62nd Prime-time Emmy Awards. August 29. Global, loose, deterministic, lasting.
- **GoogleInstantSearch**: Launch of Google Instant in United States. September 8. Global, loose, unexpected, transient.
- **HeismanTrophy**: Reggie Bush's announcement to forfeit 2005 Heisman Trophy. September 14. Local, compact, unexpected, lasting.
- **IowaStateFair**: Iowa State Fair. August 12-22. Local, loose, deterministic, lasting.
- **JewishNewYear**: Jewish New Year 5771. September 8-10. Global, compact, deterministic, transient.
- **LindsayLohanHearing**: LindsayLohan's hearing on probation revocation and verdict. September 24. Local, loose, deterministic, transient.
- **LinuxCon**: Annual convention organized by Linux Foundation. August 10-12. Global, compact, deterministic, lasting.
- **LondonTubeStrike**: London tube strike. September 6. Local, loose, deterministic, transient.
- **RichCroninDeath**: Death of singer and songwriter Rich Cronin. September 8. Local, loose, unexpected, transient.
- **ScottPilgrimRelease**: Release of movie *Scott Pilgrim vs. the World*. Aug 13. Global, loose, deterministic, lasting.
- **SESSanFrancisco**: Search Engine Strategies 2010 at San Francisco. August 16-20. Global, compact, deterministic, lasting.

[9]August and September, 2010

- **StuxnetWorm**: Confirmation of Stuxnet worm attack on Iranian nuclear program. September 24. Global, loose, unexpected, lasting.

5.1.1 Macro-level Summaries

Table 1 summarizes various statistics for all events. #tweet is the total count of tweets crawled. Following number of unique authors are the percentage of tweets having relevant url, mention, retweet and emoticon, respectively. Average subjectivity score is also reported here. The last two columns records the average size of active community over each slice and the average connectivity over each record.

5.2 Feature Vector Processing

First, all records are generated as described in Algorithm 1. Then, values of the six non-PCA content features are standardized using z-score.

We randomly sample 70% of the records for training and the rest for testing. As most information recipients did not join the community eventually, we experienced huge imbalance in terms of class labels: there are way more negative records than positive ones. To eliminate the impact of imbalanced dataset on training process, SMOTE [6] with oversampling ratio 400% is applied to positive records in the training set. After that, random under-sampling on negative records is performed for both training and testing sets to make the class distribution balanced. Finally, all numerical values are scaled to the range (-1,1), and the records are ready for evaluation.

This setting is applied to dataset of each event-oriented community with an exception that the over-sampling ratio for event *ScottPilgrimRelease* is changed to 100% for the purpose of computational efficiency.

5.3 Evaluation Settings

We run the experiments to analyze the role played by the various features and how they help us to predict whether a user will engage in the topic discussion. We use LibSVM [5] to build SVM classifiers (Gaussian RBF kernel with $\gamma = 8$ and cost $c = 32$) based on the following feature subsets to see how they perform on the prediction task. For each feature subset, the experiment is repeated five times and average accuracy rate is computed. We run the following experiment groups:

- *allFeatures* (All): contains all three feature groups.
- *onlyContent* (Con.): contains only content feature.
- *onlyAuthor* (Aut.): contains only author feature.
- *onlyCommunity* (Com.): contains only community feature.

5.4 Evaluation Results

Table 2 demonstrates the accuracy achieved by SVMs on different topics and feature sets. For each event, the highest accuracy score is in bold. Moreover, any classifier which is considered equivalently good as the highest-scoring classifier by the sign test is also in bold. We calculate the statistical significance of the improvement by performing *paired binomial sign test* on two classifiers. The smaller the p-value, the stronger evidence it is that one classifier has performance improvement over another. The p-value threshold is 0.05. Characters in the last two columns stand for U(nexpected), D(eterministic), L(oose) and C(ompact).

Our observations on experiments are listed here:

Events	All	Con.	Aut.	Com.		
DiscoveryBuildingCrisis	**77.86**	75.95	71.31	69.65	U	L
GoogleInstantSearch	**76.25**	74.92	72.23	52.60	U	L
RichCroninDeath	**90.68**	**90.96**	**90.36**	68.47	U	L
StuxnetWorm	**76.05**	**76.46**	72.05	57.51	U	L
HeismanTrophy	**76.88**	75.28	69.94	61.85	U	C
ClevelandShowPremiere	**86.11**	**85.77**	**85.65**	67.36	D	L
EmmyAwards	**77.00**	**77.39**	70.93	56.23	D	L
IowaStateFair	83.34	**84.25**	81.62	70.09	D	L
LindsayLohanHearing	**80.09**	79.30	77.22	52.57	D	L
LondonTubeStrike	**82.40**	**82.96**	80.07	56.22	D	L
ScottPilgrimRelease	**78.16**	**77.86**	75.32	59.81	D	L
JewishNewYear	**75.15**	74.14	69.16	55.63	D	C
LinuxCon	80.77	**82.17**	76.97	71.97	D	C
SESSanFrancisco	**75.50**	**76.40**	71.69	58.34	D	C

Table 2: Summary of Prediction Accuracy (%)

1) We observe performance of onlyCommunity classifiers being worst. A possible explanation for that is the latent nature of network features, which makes them difficult to be perceived by a user directly and thus have lesser effect on user engagement.

2) The onlyContent classifiers give the best performance over other single group features, especially compared to onlyCommunity classifiers. One reason for content being the dominant feature for predicting participation in a discussion is the fact that some users end up participating in a discussion based on observing the information from the public timeline, and therefore, these ad-hoc users are hard to observe via network analysis only. Moreover, content is engaging by its quality and nature (*information sharing* or *call for an action* or *crowd sourcing*). For example, link to an image or video (an evidential content) about Reggie Bush's surrender of Heisman Trophy in September, 2010 is likely to provoke lot more thoughts in a user's mind to engage in the discussion.

3) We observe comparable performance of onlyAuthor classifiers as onlyContent classifiers do for some of the topics. Here potential reason for this observation is the effective presence of influential people in the discussion group. Hence, insufficiency in content features, reflected by low average connectivity, can be compensated by author features (e.g., Rich Cronin Death).

4) Using robust statistical significance testing method, we observe for 12 out of 14 topics, allFeatures classifiers have better or equivalent performance over any single feature group classifier. In some cases (e.g. Discovery Building Crisis, a very evolving topic discussion group), the advantage is dominant, where degree of randomness in individual dimensions can be really high. Therefore, it suggest usefulness of allFeatures classifiers here.

5) We find no significant correlation between user engagement to topics and the selection of feature groups, whether the event type is *lasting* or *transient*. On the other hand, the advantage of allFeatures classifiers over other factor groups is generally stronger on the *unexpected* topics than the *deterministic* ones. Moreover, it is discovered that the performance of onlyAuthor is relatively better, explained by a closer gap to the best classifier, for *loose* events than for *compact* events.

6) The observations above suggest that we can't expect every dimension to perform well in all types of topic discussions, and hence, a strong need can be felt to study dynamics of user engagement by using the PCNA framework.

6. CONCLUSION

In this paper we present a systematic investigations into factors impacting user engagement in topic discussion on social media on his first interaction with this community. We study user engagement as problem of user participation in event-oriented community and build an effective prediction model. Evaluations on 14 Twitter event-oriented communities demonstrate that the capabilities of content, user and network features vary greatly, motivating the incorporation of all the factors. Therefore, a strong need can be felt to study dynamics of user engagement by using the PCNA framework. Moreover, we find correlations between event types and features, which can help understand user engagement in better scientific ways. Our future direction is to understand user engagement factors which keeps an user engaged to discussion topic for multiple interactions. The study will help us understand the human social dynamics on online communities.

Future research should take the following points into consideration:

- Experiments on events with more diverse characteristics for better understanding of the relation between event type and user engagement factors. Analysis of related events can help in understanding how topics around events evolve over time and shift the characteristics from one event to another.
- Sophisticated semantic analysis on user-generated content to provide content features. For example, we can resort to external knowledge base like Wikipedia to build proper context for discussion topic and then assess content quality to get better insight into impact of content features on user engagement.
- Methods to resolve user profile information's heterogeneity (e.g. missing or outdated value, adversarial content) and profile types (news, trustee etc.), and their use as people features.
- Application of the principled PCNA framework on other social networks such as Facebook, Answers.com or DBLP.
- Expanding the event-oriented model to generic framework to identify users' engagement in various co-occurring events during that timeline.

7. REFERENCES

[1] S. Asur, S. Parthasarathy, and D. Ucar. An event-based framework for characterizing the evolutionary behavior of interaction graphs. *TKDD*, 3(4):1–36, 2009.

[2] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth, and evolution. In *KDD'06*, pages 44–54. ACM, 2006.

[3] L. Backstrom and J. Leskovec. Supervised random walks: predicting and recommending links in social networks. In *WSDM'11*, pages 635–644. ACM, 2011.

[4] M. Cha, H. Haddadi, F. Benevenuto, and K. Gummadi. Measuring user influence in twitter: The million follower fallacy. In *ICWSM'04*, 2010.

[5] C. Chang and C. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[6] N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *JAIR*, 16(1):321–357, 2002.

[7] D. Cosley, D. Huttenlocher, J. Kleinberg, X. Lan, and S. Suri. Sequential influence models in social networks. In *ICWSM'10*, 2010.

[8] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In *WWW'10*, pages 591–600. ACM, 2010.

[9] E. Leicht and M. Newman. Community structure in directed networks. *Phys. Rev. Lett.*, 100(11):118703, Mar 2008.

[10] J. Leskovec, K. Lang, A. Dasgupta, and M. Mahoney. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *Internet Mathematics*, 6(1):29–123, 2009.

[11] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In *CIKM'03*, pages 556–559, 2003.

[12] C. Lin, B. Zhao, Q. Mei, and J. Han. PET: a statistical model for popular events tracking in social communities. In *KDD'10*, pages 929–938. ACM, 2010.

[13] M. Nagarajan, H. Purohit, and A. Sheth. A qualitative examination of topical tweet and retweet practices. In *ICWSM'10*, 2010.

[14] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.

[15] J. Preece. Online communities: Usability, Sociability, Theory and Methods. *Frontiers of Human-Centred Computing, Online Communities and Virtual Environments*, 2001.

[16] D. Romero, W. Galuba, S. Asur, and B. Huberman. Influence and passivity in social media. *Arxiv preprint arXiv:1008.1253*, 2010.

[17] X. Shi, J. Zhu, R. Cai, and L. Zhang. User grouping behavior in online forums. In *KDD'09*, pages 777–786. ACM, 2009.

[18] B. Suh, L. Hong, P. Pirolli, and E. Chi. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *SocialCom'10*, pages 177–184. IEEE, 2010.

[19] Y. Tsuruoka and J. Tsujii. Bidirectional inference with the easiest-first strategy for tagging sequence data. In *HLT/EMNLP'05*, pages 467–474. ACL, 2005.

[20] F. Wu and B. Huberman. Popularity, novelty and attention. In *EC'08*, pages 240–245. ACM, 2008.

[21] J. Yang and J. Leskovec. Modeling information diffusion in implicit networks. In *ICDM'10*, pages 599–608. IEEE, 2010.

[22] T. Yang, R. Jin, Y. Chi, and S. Zhu. Combining link and content for community detection: a discriminative approach. In *KDD'09*, pages 927–936. ACM, 2009.