Identification of Microblogs Prominent Users during Events by Learning Temporal Sequences of Features

Imen Bizid
L3i Lab
La Rochelle, France
imen.bizid@univ-lr.fr

Nibal Nayef
L3i Lab
La Rochelle, France
nibal.nayef@univ-lr.fr

Patrice Boursier
IUMW
Kuala Lumpur, Malaysia
patrice@iumw.edu.my

Sami Faiz
LTSIRS Lab
Tunis, Tunisia
sami.faiz@insat.rnu.tn

Antoine Doucet
L3i Lab
La Rochelle, France
antoine.doucet@univ-lr.fr

ABSTRACT

During specific real-world events, some users of microblogging platforms could provide exclusive information about those events. The identification of such prominent users depends on several factors such as the freshness and the relevance of their shared information. This work proposes a probabilistic model for the identification of prominent users in microblogs during specific events. The model is based on learning and classifying user behavior over time using Mixture of Gaussians Hidden Markov Models. A user is characterized by a temporal sequence of feature vectors describing his activities. The features computed at each time-stamp are designed to reflect both the on- and off-topic activities of users. To validate the efficacy of our proposed model, we have conducted experiments on data collected from Twitter during the Herault floods that have occurred in France. The achieved results show that learning the time-series of users' actions is better than learning just those actions without temporal information.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*selection process, retrieval models*; H.3.1 [**Information Systems**]: User/Machine Systems—*human factors, human information processing*

Keywords

Prominent microblogs users identification, Learning temporal user behavior, MoG-HMM classification

1. INTRODUCTION

Microblogging platforms represent a rich source of information indispensable to manage specific events. These platforms are seen as the perfect ground to mine relevant and exclusive information during such events. Although these microblogs such as Twitter provide many characterizing features about their content – such as the number of retweets or favorites etc. –, such features generally refer to popular users content rather than reflecting the real importance of that content. Therefore, in the context of a specific event – such as a disaster –, it is more logical to associate the relevance and quality of messages with the user's prominence in that specific event [11, 8, 15].

We define *prominent user identification* in microblogging platforms as identifying key users who provide fresh and relevant information during specific events regardless of their popularity in the platform. To the best of our knowledge, this task has never been studied in this context. However, the related problems of finding influencers have been widely explored in the state-of-the-art [10, 4]. Other works have dealt with the identification of topical authorities and domain experts [9, 13, 6]. In these related contexts, key users were identified by analyzing either the network structure using time consuming centrality algorithms such as PageRank [7, 12, 14] and HITS [1], or using the historic topical users activities independently of their temporal characteristics and off-topic activity [10, 13].

These general purpose existing approaches would give a misleading image of users behavior in real scenarios. Users are mainly evaluated according to their on-topic activity while neglecting their off-topic one. Such practice would promote official media channels toggling between several topics and which are not necessarily sharing fresh information. Moreover, users are typically represented in terms of the quantity of their produced and forwarded information independently of the temporal distribution of this information. This would give the same description for users interacting at an early stage of the event by sharing fresh information and other users posting the same information at its end.

This work is thus designed to alleviate these shortcomings. More specifically, we present the following contributions: (1) a novel representation of microblogs user behavior as a temporal sequence of features that characterize both the on- and off-topic user activities, (2) a probabilistic model for the identification of prominent users in microblogs during specific events. This model is based on learning and classifying the mentioned representation using a Mixture of Gaussians Hidden Markov Model (MoG-HMM).

The rest of this paper is organized as follows. Section 2 presents problem formulation. Section 3 discusses our proposed approach of modeling user behavior. Section 4 describes the used HMM learning model for classifying prominent users. Experimental evaluation is presented in Section 5. Finally, we conclude with directions for future work in Section 6.

2. PROBLEM FORMULATION

We formulate the problem of prominent users identification in the context of specific events as two main sub-problems. The first is representing users so that to reflect their temporal behavior during an event. Each user u has to be represented by a temporal sequence $V_u = (V_u^1, V_u^2, ..., V_u^m)$ where m is the length of the sequence describing the user behavior over m time stamps. $V_u^{(i)}$ represents user description at each time interval i, and can be any set of features characterizing the user. The second sub-problem is learning to classify users' temporal sequences of features as either belonging to prominent c_1 or non-prominent c_2 user classes. Thus we need to learn two probabilistic models H_{c_1} and H_{c_2} by training the continuous temporal sequences describing each class of users. Given these models, we need to estimate the likelihood $L(V|H_{c1})$ and $L(V|H_{c2})$ of each user sequence.

3. MODELING USERS AS TEMPORAL SEQUENCES OF ACTIVITY FEATURES

In order to model users consistently with their realistic behavior in microblogs during events, we propose a temporal sequence representation approach. The behavior of users is represented according to their observed on- and off-topic activities at different temporal stages during an event.

3.1 Temporal Sequence Representation

The time-line of an event is divided into equispaced intervals at m time stamps. During each interval, users activities are characterized by a set of features rather than a single feature as there are several types of activities in microblogs. The user activity is represented by the feature vector $V_u^{t_i}$ calculated based on $t_1, t_2, t_3, ..., t_m$ time stamps. Those features – discussed in the next subsection – describe the user behavior regarding an event (on-topic activity) and also regarding other topics (off-topic activity) during each time interval. Figure 1 illustrates – in its upper part – such a user representation.

Using this user behavior modeling approach, we provide a more enriched representation of microblogs users by considering both on- and off-topic user activeness levels over time. Such representation is capable of characterizing similarities and regularities of users behaviors.

3.2 User Activity Features

We define a set of new features inspired by the features proposed by Pal et al. [9]. For each user u and each time stamp t_m, we compute the following features by taking into account both the on-and off-topic user activities:

Topical Attachment: indicates the involvement rate of the user regarding the analyzed event by referring to the number of his original on-topic tweets ($T1_{on}$) and the number of his on-topic shared links ($T2_{on}$) adjusted by the off-topic metrics ($T1_{off}$) and ($T2_{off}$).

$$F1_u^{(t)} = (T1_{on}^{(t)} + T2_{on}^{(t)}) - (T1_{off}^{(t)} + T2_{off}^{(t)}) \qquad (1)$$

Topical Influence: estimates the value (or worthiness) of the user's original tweets regarding the event according to the number of favorites ($T3_{on}$) and the number of retweets ($R2_{on}$) of his produced tweets.

$$F2_u^{(t)} = T1_{on}^{(t)} \times log(T3_{on}^{(t)} + R2_{on}^{(t)} + 1) \qquad (2)$$

Retweeting Rate: measures the impact of the shared event-related-tweets on the user retweeting activity. This feature adjusts the number of the user's retweets ($R1_{on}$) by the number of unique users that he has retweeted ($R4_{on}$).

$$F3_u^{(t)} = R1_{on}^{(t)} * log(R4_{on}^{(t)} + 1) \qquad (3)$$

Retweeted Rate: calculates the impact of the original tweets produced by the user on the other users. This feature adjusts the number of users who have retweeted the user's on-topic tweets ($R3_{on}$) and the total number of retweets ($R2_{on}$) by the corresponding off-metrics ($R2_{off}$) and ($R3_{off}$).

$$F4_u^{(t)} = R2_{on}^{(t)} * log(R3_{on}^{(t)} + 1) - R2_{off}^{(t)} * log(R3_{off}^{(t)} + 1) \quad (4)$$

Incoming Mention Rate: measures the diversity of mentions related to the user. This feature adjusts the number of received on-topic mentions ($M1_{on}$) by the number of unique users mentioning the user ($M2_{on}$).

$$F5_u^{(t)} = M1_{on}^{(t)} * log(M2_{on}^{(t)} + 1) \qquad (5)$$

Outcoming Mention Rate: represents the mentioning activity of the current user by computing the number of his on-topic produced mentions ($M3_{on}$) adjusted by the number of unique mentioned users ($M4_{on}$).

$$F6_u^{(t)} = M3_{on}^{(t)} * log(M4_{on}^{(t)} + 1) \qquad (6)$$

Centrality Degree: promotes users having more on-topic followers and followees than off-topic ones. This feature is measured by adjusting the number of on-topic followers ($G1_{on}$) and followees ($G2_{on}$) of each user with the number of his off-topic relations ($G1_{off}$) and ($G2_{off}$).

$$F7_u^{(t)} = log(\frac{G1_{on}^{(t)} + 1}{G1_{off}^{(t)} + 2}) - log(\frac{G2_{on}^{(t)} + 1}{G2_{off}^{(t)} + 2}) \qquad (7)$$

Once these features are computed, each user u can be represented by the following feature vector at each time stamp t_m.

$$V_u^{t_i} = (F1_u^{(t)}, F2_u^{(t)}, F3_u^{(t)}, F4_u^{(t)}, F5_u^{(t)}, F6_u^{(t)}, F7_u^{(t)}) \qquad (8)$$

The set of concatenated feature vectors computed at all the time stamps represent the temporal sequence of the user behavior.

4. LEARNING TO CLASSIFY USER TEMPORAL SEQUENCES

In order to classify the time-series of feature vectors V describing each microblogs user, we train two models for prominent and non-prominent users classification using MoG HMMs. There are various types of continuous HMM: left-right, parallel left-right and ergodic. To learn our MoG HMMs, we use the ergodic model as the user activity level state at a period of time t_i can change to every other state at the period of time t_{i+1} through a single transition. Figure 1 shows a 3-state ergodic model describing how the sequence of feature vectors representing a given user can be transformed into a sequence of discrete states.

To learn the parameters for our MoG-HMM ergodic models, we use the Baum-Welch algorithm [5]. This algorithm is based on the EM algorithm to search for the maximum probability of the HMM models parameters that better fit

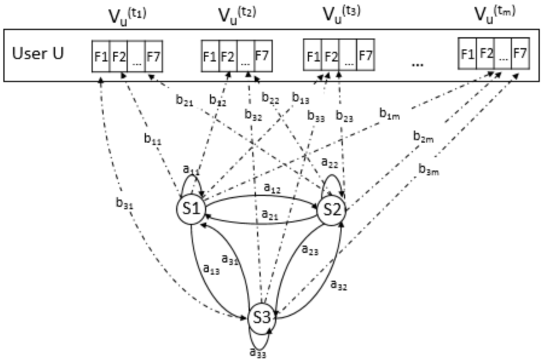

Figure 1: A time-series representation of user activities during an event using a 3-state ergodic HMM

the observed temporal users sequences in the training data.

$$H = \arg\max_{H} P(V_{training}|H) \tag{9}$$

A MoG-HMM model H is described by the quadruplet $H = \{S, \pi, A, B\}$, where $S = S_1, S_2, S_3, ..., S_k$ refers to the set of k hidden states describing levels of users activities at each period of time t_i. The state of a user at time t can expressed by $(X_t \in S)_{1 \le t \le m}$. π denotes the initial probability of the different states. A is the state transition probability matrix to change from state S_i to S_j, $A = a_{ij}$ where $a_{ij} = P(X_{t+1} = S_j | X_t = S_i)_{1 \le i,j \le k}$. B refers to the continuous output probability matrix where the probability $B = b_i(V^t)$ represents the probability of observing a feature vector V^t from a state S_i, where $b_i(t) = P(V^t|X_t) = (S_i)_{1 \le i \le k}$.

The transformation of these feature vectors into discrete states is processed by the construction of a continuous observation probability density function (PDF) matrix B. This matrix is represented as a mixture of Gaussians in order to associate the sequence of a user's feature vectors into the different finite states using equation 10.

$$b_i(V^t) = \sum_{k=1}^{M} c_{ik} \mathcal{N}[V^t, \mu_{ik}, W_{ik}] \tag{10}$$

where c_{ik} is the mixture weight, $\mathcal{N}$ is the normal density, μ_{ik} is the mean vector and W_{ik} is the covariance matrix for the k^{th} mixture component in state S_i.

Once the models parameters H_{c1} and H_{c2} are set through training, we can compute the probability of any microblogs user to belong to each class $P(V_u|H_{c1})$ and $P(V_u|H_{c2})$ given the two learned models. These probabilities are obtained using the forward-backward algorithm [2]. If the model H_{c1} gives a higher probability to a represented user compared to $P(V_u|H_{c2})$, then this user is classified as prominent.

5. PERFORMANCE EVALUATION

5.1 Dataset

To conduct experimental performance evaluation on real data, we have tracked the Twitter users who have shared at least one on-topic information about the floods that have occurred from 29^{th} to 30^{th} September 2014 in the Herault area, situated in the south of France using ou multi-agent system MASIR [3]. 3332 users have been tracked during the two days of the disaster.

To create the ground-truth, we conducted a subjective user study in order to label each user in the dataset with **c1** and **c2** indicating respectively whether the user is prominent or not. The tracked users were evaluated according to the relevance and exclusivity of their tweets during the disaster. According to this study, 90 users have been classified as $c1$, and 3242 users have been classified as $c2$.

5.2 Evaluation Set-up and Metrics

For experimental set-up, we randomly sampled 60% of both prominent and non prominent labeled users datasets as training data for building the H_{c1} and H_{c2} classifiers, and the remaining 40% as test data. Features characterizing user behavior were extracted sequentially at each time interval of 90 minutes from the beginning of the event to its end. Thus each user is represented by a sequence of 32 feature vectors. We have also extracted features using different interval lengths.

Following the standard evaluation criteria used in the context of prominent users identification, we use **precision, recall** and **F1-score** measures to evaluate the performance of our approach. In order to choose the coherent parameters for the representation of microblogs users behavior through H_{c1} and H_{c2} models, we have tested different values of "number of states" N_s (from 1 to 6) and "number of multivariate Gaussian" N_G (from 1 to 5) with the training dataset. The experimental results are shown in Table 1. The best results are obtained when N_s = 3 states and N_G = 1, yielding F1-score value to 64%.

Table 1: Prominent users identification performance for different N_s and N_G in terms of F1-score.

N_s/N_G	1	2	3	4	5
1	0.5201	0.5615	0.6035	0.5643	0.6039
2	0.5500	0.5755	0.01	0.01	0.01
3	**0.6455**	0.01	0.01	0.1337	0.2026
4	0.5855	0.01	0.1100	0.01	0.1020
5	0.6156	0.01	0.01	0.01	0.01
6	0.01	0.4700	0.01	0.01	0.01

5.3 Experimental Results

5.3.1 *Importance of Time-series Representation*

To demonstrate the effectiveness of our temporal sequence representation approach for the identification of prominent users, we test the performance of our model by decreasing the length of the feature vectors sequence m(from 32 to 2) (e.g. m=2 users activities features are recorded at each 720 minutes). In other words, users behavior are represented during longer periods of time. Figure 2 shows how the sequence length variation affects the performance of our model. According to these results, we find that larger sequence length characterizing detailed users activities over time works significantly better than smaller ones. Thus, user behavior is better characterized with more time stamps.

5.3.2 *Comparison with Different Models*

We compared our HMM temporal sequence classification model with two baseline models:

Baseline 1: refers to an SVM model learned using our

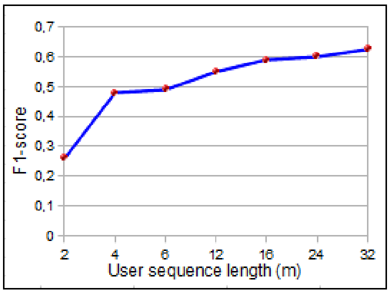

Figure 2: Time-series sequence length effect on the model performance

Figure 3: The different ergodic MoG-HMM models appropriated to each phase

proposed features and a classic single vector representation of the user behavior during the event.

Baseline 2: refers to an SVM model learned using Pal et al. features [9] and a single vector user representation.

Figure 3 shows the results of comparing the three different models, where our proposed model significantly outperforms the models which are trained on users activities in the whole event duration (i.e independently of the activity distribution over time). We also note that our features are more effective in the context of prominent users identification in specific events than Pal et al. [9] features which are based on analyzing only the on-topic user activity.

6. CONCLUSION AND FUTURE WORK

This paper has presented a novel microblogs users representation according to their behavior over time in specific events. This representation has been used for building and training a MoG-HMM model to identify prominent users. In our model, users are characterized by a temporal sequence composed of feature vectors recorded in different periods of time during the event. These features characterize the on-topic and off-topic users activities at each time interval. Our experiments show that a longer length of the sequence characterizing the user behavior over time, produces a better identification model. We have additionally compared the performance of our model to traditional machine learning SVM models using our features and state-of-the-art features. The achieved results show the importance of learning time sequences of users activities as compared to learning those activities independently of their temporal characteristics.

For future work, we aim to predict prominent users at different event phases. We would like also to analyze the correlation between temporal activities of prominent users and the time of occurrence of the trending sub-events.

7. REFERENCES

[1] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. Finding high-quality content in social media. In *WSDM*, pages 183–194, 2008.

[2] L. E. Baum and J. A. Eagon. An inequality with applications to statistical estimation for probabilistic functions of markov processes and to a model for ecology. *Bull. Amer. Math. Soc.*, 73(3):360–363, 1967.

[3] I. Bizid, P. G. Boursier, J. Morcos, and S. Faiz. Masir : A multi-agent system for real-time information retrieval from microblogs during unexpected events. KES-AMSTA-15, pages 1–8, 2015.

[4] W. Chen, Y. Wang, and S. Yang. Efficient influence maximization in social networks. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 199–208, 2009.

[5] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.

[6] S. Ghosh, N. Sharma, F. Benevenuto, N. Ganguly, and K. Gummadi. Cognos: Crowdsourcing search for topic experts in microblogs. In *SIGIR*, pages 575–590, 2012.

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

[8] Q. V. Liao, C. Wagner, P. Pirolli, and W.-T. Fu. Understanding experts' and novices' expertise judgment of twitter users. In *CHI*, pages 2461–2464, New York, NY, USA, 2012.

[9] A. Pal and S. Counts. Identifying topical authorities in microblogs. In *WSDM*, pages 45–54, 2011.

[10] D. M. Romero, W. Galuba, S. Asur, and B. A. Huberman. Influence and passivity in social media. In *WWW*, pages 113–114, 2011.

[11] C. Wagner, V. Liao, P. Pirolli, L. Nelson, and M. Strohmaier. It's not in their tweets: Modeling topical expertise of twitter users. In *SocialCom*, pages 91–100, 2012.

[12] J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: Finding topic-sensitive influential twitterers. In *WSDM*, pages 261–270, 2010.

[13] S. Xianlei, Z. Chunhong, and J. Yang. Finding domain experts in microblogs. In *WEBIST*, 2014.

[14] M. Zhang, C. Sun, and W. Liu. Identifying influential users of micro-blogging services: A dynamic action-based network approach. In *PACIS*, page 223, 2011.

[15] G. Zhou, S. Lai, K. Liu, and J. Zhao. Topic-sensitive probabilistic model for expert finding in question answer communities. In *CIKM*, pages 1662–1666, 2012.