Prominent Users Detection during Specific Events by Learning On- and Off-topic Features of User Activities

Imen Bizid, Nibal Nayef, Patrice Boursier, Sami Faiz, Jacques Morcos

Prominent Users Detection during Specific Events by Learning On- and Off-topic Features of User Activities

Imen Bizid*†, Nibal Nayef*, Patrice Boursier*‡†, Sami Faiz†, and Jacques Morcos*
*L3i, University of La Rochelle, Avenue Michel Crépeau, 17042, La Rochelle, France
Emails: {imen.bizid, nibal.nayef, patrice.boursier, jacques.morcos}@univ-lr.fr
†LTSIRS, University of Tunis, 1005, Tunis, Tunisie
Emails: {sami.faiz}@insat.rnu.tn
‡IUMW, City Campus, Jalan Tun Ismail, 50480 Kuala Lumpur, Malaysia
Emails: {patrice}@iumw.edu.my

***Abstract*—Microblogs such as Twitter are characterized by the richness and recency of information shared by their users during major events. However, it is very challenging to automatically mine for information or for users sharing certain information due to the huge variety of unstructured stream of data shared in such microblogs. This work proposes a ranking and classification model for identifying users sharing useful information during a specified event. The model is based on a novel set of features that can be computed in real time. These features are designed such that they take into account both the on and off-topic activities of a user. Once users are characterized by a feature vector, supervised machine learning tool is trained to classify users as either prominent or not. Our model has been tested on data shared during a flooding disaster event and performed very well. The achieved results show the effectiveness of the proposed model for both the classification and ranking of prominent users in such events, and also the importance of the adjustment of the on-topic features by the off-topic ones when describing users' activities.**

Keywords—Key user identification, On- and off-topic features, Events in Twitter, SVM Classification and Ranking, Real-time feature computation.

I. INTRODUCTION

Microblogging platforms offer the services of convenient access to and sharing of fresh data on any topic. Shared data is usually limited to a specified number of characters, for example, tweets are limited to 140 characters. This results in a huge stream of unstructured data, which makes information retrieval within such data very challenging, specially when having to perform such tasks in real time.

Having the aforementioned particularities of microblogs in mind, many research works have associated the relevance and the quality of the shared messages in microblogs with the user's prominence in terms of the network and the targeted topic but not in terms of messages' content [1], [2].

Those work have focused on the identification of influencers [3], [4] and domain experts [5], [6] in microblogs. The identification of the latter type of users is usually based either on their centrality and popularity in the network or on their frequent networking activities domains. From another aspect, domain experts are generally identified by analyzing their historic information regarding a topic of interest. However, the prominence of microblogs users during a *specific event* cannot be evaluated according to the user's centrality or prior activities in the network.

The features characterizing prominent users should reflect the nature of a user's behavior during such events. Consider the case of a disaster where alerts and emergency states are being rapidly updated over time. The prominent users – whom we are interested in – would focus their attention and communication mostly on the topical information related to the disaster. These users are not necessarily experts in disasters, they may be ordinary microblogs users geolocated in the disaster area and who are sharing what they are seeing and experiencing. Hence, they would share a lot of exclusive information. While users such as media channels toggling between several topics would share many event-related information which are already diffused in the network. Thus, features based on traditional metrics such as the number of shared on-topic tweets cannot be directly used to identify user's prominence. Therefore, identification models which are based on traditional metrics would be sensitive to users interested in several topics and sharing outdated information.

In this paper, we propose a model that alleviates these shortcomings first by designing a set of metrics which represent Twitter users interested in a specific event. These metrics characterize each active user by considering both the on- and off-topic activities of the user during the event. Based on those metrics, our list of features promote users who focus only on the event under consideration, and penalize those who are toggling among several topics. Using these representative features, we use supervised learning to train an SVM model to identify the most prominent users in real time during an event.

The rest of this paper is organized as follows: Section II reviews related work. Section III presents the set of our proposed features used to model microblogs' users. Section IV summarizes the classification and ranking approach employed to identify and detect the best prominent users. Section V presents the experiments and results obtained by our model. Section VI concludes the paper and discusses future work.

II. RELATED WORK

To the best of our knowledge, the issue of prominent users' identification has never been explored in the context of specific events. However, there have been several attempts proposing new measures identifying influential users and domain experts in microblogs and specially Twitter [7], [8].

As the identification of social networks influencers problem has been potentially associated with the global issue of central users finding in any kind of network, most of the proposed approaches identifying social media influencers are based on standard centrality measures such as PageRank [7] and HITS [9]. These measures have proved their potential to identify influencers according to their social position in microblogs. However, they are still sensitive to celebrities and news channels that have a large number of followers. Moreover, microblogs are richer than simple links relating users.

Therefore, many approaches have adapted PageRank and Hits algorithms to the specificities of microblogs (e.g. the number of shared tweets, number of mentions and number of retweets etc.) [10], [11]. TwitterRank [12] measures the user's influence based on both the topical similarity between users and their link structure. IP-influence [3] weights the edges by taking into account microblogs features related to user's passivity and influence rates and measures the user's influence using HITS algorithm. This approach is similar to KHYRank [13] which uses the number of retweets and mentions to weight the graph. While these adapted PageRank and Hits algorithms have yielded better results for influencers identification, their iterative process is still time consuming and unsuitable for the identification of prominent users during an event.

Few prior studies have explored the detection of topical authorities or domain experts in real-time scenarios [14], [5], [15]. Cognos system [15] is built for finding domain experts in Twitter by mining Twitter lists created by individual users to include experts on topics. Xianlei et al. [5] proposed a Gradient Boosted Decision Tree to identify domain experts in Sina Microblog over state-of-the-art [14] and new linguistic features. Pal et al. [14] proposed a clustering and ranking procedure based on a set of features including nodal and topical metrics characterizing microblog user's activities.

All these studies have focused on measuring the attachment of microblogs users only for specific analyzed topics while neglecting their activities on other topics. These approaches are sensitive to active users interested in several topics and who are not necessarily sharing updated or detailed information about the topic of interest. Whereas in this paper, we propose a new model built based on features that describe the user's activities on both the analyzed topic and the other ones. This model is capable of identifying prominent users in specific and unexpected events such as natural or human disasters.

III. Describing Users by their On- and Off-topic Activity Features

In order to build a classification model for identifying prominent users interacting about an event, we propose a set of new features to describe and characterize users and their behavior during that event.
These features are composed of a set of metrics computed according to both the on- and off-topic activities of microblogs users during the event.
On-topic: An activity is considered on-topic when it contains a subset of a list of keywords and hashtags which are defined to describe the event under consideration.
Off-topic: an off-topic activity refers to any activity that was not recorded as an on-topic one.

Nevertheless, if a captured tweet referring to the event includes some keywords reflecting non-serious or non-valuable contents (e.g. advertising or joke words and symbols such as sale, rent, pub, lol and so on), it will be directly recorded as an off-topic one. Our rationale behind the extraction of on-topic and off-topic activities is based on penalizing users who are toggling among several topics, and who may share outdated information.

A. *Characterizing of Users' Activities*

Users' activities in Twitter falls into three types: *original tweets* are tweets originally expressed by a user, *retweets* are tweets already shared by someone else in the network and forwarded later by a user and *mentions* are tweets directed to particular users mentioned using the @ symbol.

Table I shows the complete list of set of our proposed and the state-of-the art metrics used in this work to describe the different user's activities during a specific event.

TABLE I. List of metrics that describe users and their activities in Twitter. **(new)** denotes the new proposed metrics, **On(+)** and **Off(+)** refer to the recorded metrics for both on- and off-topic activities of users

	On	Off
Original tweets		
T1: Number of original tweets [14], [5]	+	+
T2: Number of links shared [16]	+	+
T3: Number of keyword and hashtags [14]	+	-
T4: Number of favorites of original tweets (new)	+	+
T5: Number of tweets geo-located in the event area (new)	+	-
Retweets		
R1: Number of retweets of other's tweets [17], [5]	+	+
R2: Number of unique users retweeted by the user (new)	+	+
R3: Number of retweets of author's tweets (new)	+	+
R4: Number of unique users who retweeted author's tweets [17]	+	+
Mentions		
M1 : Number of mentions of other users by the author [18], [14]	+	+
M2 : Number of unique users mentioned by the author [18], [14]	+	+
M3 : Number of mentions by others of the author [18], [14]	+	+
M4 : Number of unique users mentioning the author [18], [14]	+	+
Graph		
G1: Number of active followers [7]	+	+
G2: Number of active followees [7]	+	+

B. *Building a User-level Feature Vector*

Inspired by the features presented in Pal et al. [14], we propose a new list of features to represent users by adjusting their on-topic activities by their off-topic ones. The features are based on the list of metrics presented in Table I. In the following, we explain those features in detail:
Topical Strength: estimates the value (or worthiness) of the author's topical tweets with respect to the off-topic ones. $F1$ promotes users that have collected more favorite points regarding their on-topical tweets than off-topical ones.

$$F1 = \frac{T4_{on}}{T4_{off} + 1} \tag{1}$$

Topical Attachment: indicates the involvement rate of the user regarding the analyzed topic by referring to the number of his original on-topic tweets adjusted by the off-topic ones. The more a user produces on-topical tweets compared to off-topical ones, the higher his Topical Attachment score is.

$$F2 = \frac{T1_{on} + T2_{on}}{T1_{off} + T2_{off} + 1} \tag{2}$$

Retweeting Rate: measures the impact of the original tweets shared by other users on the user's topical activities. This measure is adjusted by the retweeting activity of the user regarding off-topic original tweets.

$$F3 = R1_{on} * log(R2_{on} + 1) - R1_{off} * log(R2_{off} + 1) \quad (3)$$

Retweeted Rate: calculates the impact of the topical original tweets produced by the author on other network users. This feature is adjusted by the user's influence rate on other topics.

$$F4 = R3_{on} * log(R4_{on} + 1) - R3_{off} * log(R4_{off} + 1) \quad (4)$$

Incoming Mention Rate: measures the diversity of mentions that the user has received regarding the specific topic. This measure is adjusted by the flow rate of off-topic mentions intended to the user.

$$F5 = M3_{on} * log(M4_{on} + 1) - M3_{off} * log(M4_{off} + 1) \quad (5)$$

Outcoming Mention Rate: promotes users producing many on-topic mentions intended to several users on the one hand and penalizes users having more off-topic mentions addressed to different users than on-topic ones on the second hand.

$$F6 = M1_{on} * log(M2_{on} + 1) - M1_{off} * log(M2_{off} + 1) \quad (6)$$

Centrality Degree: adjusts the number of on-topic followers and followees of each user with the number of his off-topic relations. This feature promotes users having more on-topic followers than off-topic ones.

$$F7 = log(\frac{G1_{on} + 1}{G1_{off} + 2}) - log(\frac{G2_{on} + 1}{G2_{off} + 2}) \quad (7)$$

These hand-crafted features combine different metrics which can be extracted in real time from users' time-line using Twitter APIs. Using these features, we offer a better representation of users by combining different metrics which have significant relations among themselves.

Thus by computing the above described features, we model each user by the following feature vector composed of seven features describing his on- and off-topic activities.

$$x_i = (F1_i, F2_i, F3_i, F4_i, F5_i, F6_i, F7_i, T5_i) \quad (8)$$

IV. Classification and Ranking of Prominent Users

To identify the prominent users from within the huge number of users that may be interacting during a specific event, we model this problem into a binary classification problem (i.e. 1 for prominent users or -1 for non-prominent ones). We use a supervised learning method in order to build our classification model. The role of the classification is to detect the prominent users and reject the others. Then, in the ranking stage, we mainly focus on identifying the top prominent users regarding the specific event or topic under consideration.

V. Performance Evaluation

A. Dataset

To conduct experimental performance evaluation on real data, we collected most of the tweets shared during the floods that have occurred from 29^{th} to 30^{th} September 2014 in the Herault area, situated in the south of France. Data collection was processed using our multi-agent System called MASIR [19]. At the lowest level, the system detects the different users who have shared at least one on-topic tweet (i.e. talking about the floods) during the analyzed period. On-topic tweets are detected using the hashtags and keywords which were employed by Twitter users to share information about the disaster. The system then crawls all the on-topic and off-topic tweets shared by the detected users during the event. We collected 60195 tweets composed of on- and off-topic tweets shared by 3332 users during the two days of the disaster.

Ground Truth: For the purposes of training and evaluation, we conducted a subjective study for labeling each user in the dataset as one of the two classes: C1 for prominent users, or C2 for non-prominent ones according to the relevance and freshness of his tweets' content. In this study, multiple persons have manually labeled each user in the dataset. Then, they assigned a score indicating the prominence degree of each prominent user. The first label is to be used for training and evaluating our classification model, while the second label (prominence degree score) is to be used to evaluate our user ranking model.

B. Experimental Setup

The dataset described in the previous Subsection has been divided into training and test sets using two different partitions as described in Table II. Furthermore, in order to avoid any bias in experiments, we have applied the principle of 3-fold cross validation for both partitions 1 and 2.

TABLE II. THE DIFFERENT PARTITIONS OF DATA USED IN THE TRAINING AND TEST PHASES.

	Partition 1		Partition 2	
	Training$_1$ (60%)	Test$_1$ (40%)	Training$_1$ (80%)	Test$_1$ (20%)
C1	54	36	72	18
C2	1945	1297	2593	649

C. Experiments and Results

1) Performance of our Classification Model: We compared our proposed classification model with several baselines and state-of-the-art methods as described in the following: *Our model*: using our proposed features (Subsection III-B) computed using both the on and off-topic metrics. *B1*: this model also uses the features described Subsection III-B, but without any adjustment of on-topic metrics with off-topic ones. *B2*: this model uses all the on-topic metrics presented in Table I. *B3*: this model uses the features proposed by Pal et al. [14].

Figure 1 shows the precision and recall results of our classification model to identify the prominent users compared to the other baselines. We note that the results of our classification model are significantly higher than the other baseline models. According to the recall results of the two partitions, we observe that our model detects most of the true prominent users, and achieves between 8% to 20% higher recall than the baseline methods. Additionally, we note that the precision of the different models for class C1 is well under 50%. However, this result remains important, as the different classification models have rejected most of the non-prominent users and performed worse than our model. Overall, through these experiments, we establish that our model outperforms

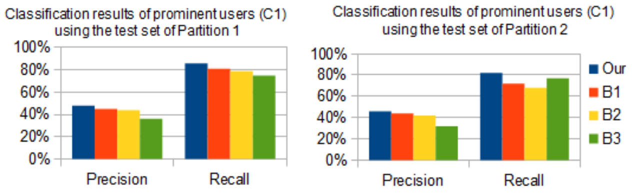

Fig. 1. Classification results of our proposed method for class **C1** according to precision and recall metrics.

other baseline methods which use only on-topic features to represent the user importance. Hence, we demonstrate that adjusting on-topical metrics with off-topical ones improves the classification results.

2) Performance of our Ranking Model: According to the classification results, our model has identified most of the true prominent users in the different partitions. However, it misclassified a small number of non-prominent users. Therefore, we need to evaluate the efficiency of our ranking model in the detection of the top prominent users. We compare our model with the following two baseline methods: *Our model:* we use our proposed features (Subsection III-B) computed using both the on and off-topic metrics. *Baseline 1:* this model uses the features proposed by Pal et al. [14]. *Baseline 2:* this model uses the PageRank algorithm in order to measure the score of each user according to its centrality in the network. Thus, we have constructed a network relying on the different users who have shared at least one on-topic tweet about the event. We have ranked the set of users identified in class C1 using the different ranking baseline models. Then, we picked the top 15 users detected by each baseline. The precision of the ratings accorded by each baseline are computed by counting the number of correctly detected users in the top 15 with respect to our ground truth. The results of these experiments are illustrated in Figure 2. According to these results, we observe that our model achieves the highest precision compared to other baseline methods, with a precision of 86% in Partition 2. Therefore, our designed high level features outperform the state-of-the-art features for both the identification of prominent users and the detection of the top ones. Moreover, we note that PageRank algorithm achieves the worst results compared to the models constructed using machine learning techniques.

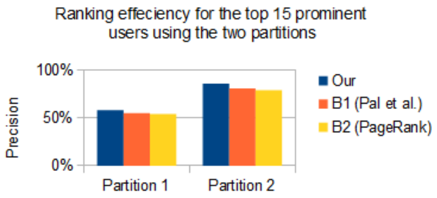

Fig. 2. Performance evaluation of our model according to Ranking precision of users classified in class C1 (prominent users).

VI. Conclusion and Future Work

This paper has presented a classification and ranking model to identify prominent users in a specific topic or event. Through the conducted experiments, we have shown that models learned according to high or low level features computed from both on- and off-topic metrics outperform other models that are based only on on-topic features. Despite the challenges posed by the nature of our real data, we have shown how the used supervised learning algorithm (SVM) can still be effective with appropriate features and using different weights for imbalanced data classes.

For future work, we aim to explore the different features proposed in the literature to model microblogs' users in order to select the most representative features for prominent users. In addition, we plan to investigate the use of other machine learning models to predict the top important users.

References

[1] C. Wagner, V. Liao, P. Pirolli, L. Nelson, and M. Strohmaier, "It's not in their tweets: Modeling topical expertise of twitter users," ser. =SocialCom '12, Sept 2012, pp. 91–100.

[2] Q. V. Liao, C. Wagner, P. Pirolli, and W.-T. Fu, "Understanding experts' and novices' expertise judgment of twitter users," ser. CHI '12. New York, NY, USA: ACM, 2012, pp. 2461–2464.

[3] D. M. Romero, W. Galuba, S. Asur, and B. A. Huberman, "Influence and passivity in social media," ser. WWW '11, 2011, pp. 113–114.

[4] W. Chen, Y. Wang, and S. Yang, "Efficient influence maximization in social networks," ser. KDD '09. New York, NY, USA: ACM, 2009, pp. 199–208.

[5] S. Xianlei, Z. Chunhong, and J. Yang, "Finding domain experts in microblogs," ser. WEBIST'14, 2014.

[6] A. Bozzon, M. Brambilla, S. Ceri, M. Silvestri, and G. Vesci, "Choosing the right crowd: Expert finding in social networks," ser. EDBT '13. New York, NY, USA: ACM, 2013, pp. 637–648.

[7] H. Kwak, C. Lee, H. Park, and S. Moon, "What is twitter, a social network or a news media?" ser. WWW '10. New York, NY, USA: ACM, 2010, pp. 591–600.

[8] L. C. Freeman, D. Roeder, and R. R. Mulholland, "Centrality in social networks: ii. experimental results," *Social Networks*, vol. 2, no. 2, pp. 119 – 141, 19791980.

[9] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne, "Finding high-quality content in social media," ser. WSDM '08. New York, NY, USA: ACM, 2008, pp. 183–194.

[10] R. Cappelletti and N. Sastry, "Iarank: Ranking users on twitter in near real-time, based on their information amplification potential," ser. SOCIALINFORMATICS '12. Washington, DC, USA: IEEE Computer Society, 2012, pp. 70–77.

[11] A. Silva, S. Guimarães, W. Meira, Jr., and M. Zaki, "Profilerank: Finding relevant content and influential users based on information diffusion," ser. SNAKDD '13. New York, NY, USA: ACM, 2013, pp. 2:1–2:9.

[12] J. Weng, E.-P. Lim, J. Jiang, and Q. He, "Twitterrank: Finding topic-sensitive influential twitterers," ser. WSDM '10. New York, NY, USA: ACM, 2010, pp. 261–270.

[13] M. Zhang, C. Sun, and W. Liu, "Identifying influential users of micro-blogging services: A dynamic action-based network approach." in *PACIS'11*. Queensland University of Technology, 2011, p. 223.

[14] A. Pal and S. Counts, "Identifying topical authorities in microblogs," ser. WSDM '11. New York, NY, USA: ACM, 2011, pp. 45–54.

[15] S. Ghosh, N. Sharma, F. Benevenuto, N. Ganguly, and K. Gummadi, "Cognos: Crowdsourcing search for topic experts in microblogs," ser. SIGIR '12, 2012, pp. 575–590.

[16] A. Java, Pranam, T. Finin, and T. Oates, "Modeling the spread of influence on the blogosphere," in *www*, 2006.

[17] D. Boyd, S. Golder, and G. Lotan, "Tweet, tweet, retweet: Conversational aspects of retweeting on twitter," ser. HICSS' 09, Jan 2010, pp. 1–10.

[18] C. Honey and S. Herring, "Beyond microblogging: Conversation and collaboration via twitter," ser. HICSS '09, Jan 2009, pp. 1–10.

[19] 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," ser. KES-AMSTA-15, 2015, pp. 1–8.