Information Systems 000 (2018) 1–16

Contents lists available at ScienceDirect

Information Systems

journal homepage: www.elsevier.com/locate/is

Detecting prominent microblog users over crisis events phases

Imen Bizid*, Nibal Nayef, Patrice Boursier, Antoine Doucet

L3i, University of La Rochelle, France

ARTICLE INFO

Article history:
Received 14 April 2017
Revised 11 December 2017
Accepted 15 December 2017
Available online xxx

Keywords:
Information retrieval from microblogs
Prominent users prediction
User behavior modeling
Phase-temporal representation

ABSTRACT

During crisis events such as disasters, the need for real-time information retrieval (IR) from microblogs becomes essential. However, the huge amount and the variety of the shared information in real time during such events over-complicates this task. Unlike existing IR approaches based on content analysis, we propose to tackle this problem by using user-centric IR approaches with identifying and tracking prominent microblog users who are susceptible to share relevant and exclusive information at an early stage of each analyzed event phase. This approach ensures real-time access to the valuable microblogs information required by the emergency teams. In this approach, we propose a phase-aware probabilistic model for predicting and ranking prominent microblog users over time according to their behavior using Mixture of Gaussians Hidden Markov Models (MoG-HMM). The model utilizes a new user representation which takes into account both the user and the event specificities over time. This user representation comprises the following new aspects (1) Modeling microblog users behavior evolution by considering the different event phases (2) Characterizing users activity over time through a temporal sequence representation (3) Time-series-based selection of the most discriminative features (4) prominent users prediction using probabilistic phase-aware models learned *a priori*. We have conducted experiments during flooding events: we trained our identification models using a dataset relative to the “Alpes-Maritimes floods” and we tested its identification performance using a new dataset relative to another flooding disaster “Herault floods”. The achieved results show that our model significantly outperforms phase-unaware models and identifies most of the prominent users at an early stage of each event phase.

1. Introduction

The effectiveness and ease-of-use of supported microblogging platforms – such as Twitter – have revolutionized the communication habits in our society. Any user can quickly and conveniently post and get information on the latest news. During crisis events, the amount of communicated information in such platforms increases significantly. This makes information retrieval more challenging. Most of the shared tweets during these events are non-valuable, redundant, outdated or incredible. Moreover, this shared data is generally expressed in several languages and various formats (i.e. texts, images, links and videos). Thus, content-based retrieval approaches are not well suited for this task as they are time consuming.

This information retrieval problem has been addressed in the literature by associating the quality of tweets with the prominence of their authors in a specific topic or event [1,2]. In the context of this article, we define *prominent users* as microblog users who are susceptible to share relevant and exclusive information during crisis events regardless of their popularity and their domain of expertise in the platform. To the best of our knowledge such users category has never been targeted in the literature. However, there have been several works targeting other categories of important authors known as domain experts or topical authorities.

The detection of these categories has gained a wide interest in the literature. However, the detection techniques proposed for these categories are not suitable to identify prominent users targeted in this paper. Prominent users in the context of crisis events cannot be systematically considered as domain experts or topical authorities. Most of prominent users refer to *ordinary* users who may provide their testimony based on what they experience in the region of a crisis.

Targeted key users in the literature are generally identified using ranking techniques based on either a graph-based user modeling approach or a vector-based user modeling approach. Graph-based approaches are sensitive to popular microblog users who have a large number of connections, such as celebrities and news outlet channels.

* Corresponding author.
E-mail addresses: imen.bizid@univ-lr.fr (I. Bizid), nibal.nayef@univ-lr.fr (N. Nayef), patrice.boursier@univ-lr.fr (P. Boursier), antoine.doucet@univ-lr.fr (A. Doucet).

https://doi.org/10.1016/j.is.2017.12.004
0306-4379/© 2017 Elsevier Ltd. All rights reserved.

Vector-based user modeling approaches have thus been proposed to deal with this problem. Such approaches describe users by a vector of features reflecting the overall tweeting activity on each user based on textual, microblogging and social network structure features. However, such vector-based user modeling approach does neither realistically nor accurately represent the evolution of user behavior over time. This yields weaker performance of detection and ranking algorithms which learn to distinguish behavioral differences among different users.

Characterizing users without considering the temporal distribution of their activities over event phases would not reveal the real user behavior. This is due to the following: (1) Quantitative characterization of users: Practically, such characterization would promote users sharing much information about an event even if this information is irrelevant or outdated. (2) Uniform user characterization over the event duration (from the beginning of an event until its end): Realistically, the behavior of users may differ according to the evolution of the event. (3) Overall user prominence evaluation over the event duration: Such strategy would fail to discover true prominent users who were active in only one – however important – event phase, because their activity statistics are lower compared to other users who were active in prior phases.

Moreover, the problem of key users prediction has never been tackled in the literature. Most of the proposed identification models have been modeled and experimented to classify or/and rank such users by the end of an event and not over time. The challenge behind our proposed model is to predict such prominent users at an early stage of each event phase in order to track these users and get access to the relevant information they are sharing.

This work alleviates these shortcomings by proposing a new user modeling and prediction approach considering: -Event evolution over time, and -User behavioral change over event phases and over time of each phase. Crisis events – specially natural disasters – are usually described in terms of "*phases*" having their specific goals, characteristics and experts. Each phase influences users' behavior differently according to their interest and involvement in that phase. This proposed user modeling approach is implemented within prominent users prediction and ranking model learned using data from prior crisis events. This model overcomes the problem of time-consuming information retrieval techniques by considering features which can be computed in real time and learning *a priori* the identification models adapted to each category of crisis events. Through our experiments, we have trained a model adapted for prominent microblog users identification during flooding events.

The rest of this paper describes the integration of these ideas for prominent users identification during crisis events. In Section 3, we describe our phase-aware user behavior modeling approach. We list the different extracted user features used by the feature selection process to characterize user behavior at each event phase in Section 4. Our temporal phase-aware probabilistic model for the classification and ranking of microblog user's behavior is detailed in Section 5. The evaluation set-up is described in Section 6. Experimental evaluation is presented in Section 7. Finally, we present the discussion and conclusions along with directions for future work in Sections 8 and 9.

2. Related work

To the best of our knowledge, the issue of prominent users identification has not been explored in depth in the context of crisis events. However, there have been several works which proposed models to identify other categories of key users such as microblog influential users, topical authorities and domain experts in a more general context [2,3]. Such models have mainly focused on proposing a user modeling approach that is able to highlight the differences between key and non-key users on specific topics. Through that user characterization, machine learning or ranking algorithms are generally explored to learn or identify similar users behaviors. These state-of-the-art user modeling approaches fall into two categories -A graph-based user modeling approach describing user interaction in the network [4–6] or a vector-based user modeling based on a list of descriptive user behavior features [1,2,7].

The graph-based-user modeling approach represents users behavior by a graph composed of nodes and edges denoting respectively users and any nature of relation that may link them. Such representation is generally adopted for both influencers [3,8] and domain experts detection [5]. The IP-influence model [3] – which identifies influencers – defined edges as pairwise influence and passivity according to the retweeting activity of users. TwitterRank [5] identified domain experts using the PageRank algorithm ranking users according to their position on Twitter graph constructed according to the tweeting activity of users. Such user representation has been criticized as it makes the identification process sensitive to popular users who are not necessarily prominent [1].

The vector-based user characterization has been proposed as a new alternative to address this sensitivity. This user characterization approach was firstly introduced by Pal and Counts [1] in the context of domain experts identification. They represented by a single vector composed of 15 features describing the user tweeting activity in order to cluster and rank each user according to his/her expertise. Similarly, Xianlei et al. [2] employed this same user characterization by referring to linguistic, user activity and profile features in order to classify them using a machine leaning algorithm. Ghosh et al. [7] represented users by a topic vector composed of different weighted terms extracted from the Twitter lists. Through this representation, users are ranked by computing the topical similarity scores between the different vectors.

While most of those vector-based models which identify domain experts have been applied in topics referring to events such as "The world cup", these models remain unsuitable for the context of crisis events. Firstly, prominent users – in crisis events – are not necessarily domain experts, they may be ordinary users who are implicated involuntarily in a particular disaster which has occurred in their region. Thus, such users cannot be detected *a priori* using Twitter lists [7]. Second, characterizing users uniformly and quantitatively during the whole event using such representation would not reflect the real user behavior [9]. The user behavior and interest change over time according to the evolution of the event. Finally, the user's prominence may not be associated with the whole event, users may be prominent only in one particular phase.

The present contribution addresses these limitations. In a previous work [9], we have presented a new user characterization approach consisting of representing users by a sequence of feature vectors extracted over time independently of the event characteristics. In this paper, we propose a complete user characterization considering both the user behavior and the event evolution over time in order to predict prominent users in real time. We also tackle the problem of prominent users identification in terms of prediction, not classification. In other words, we focus on learning a model which is able to predict prominent users over time and not by the end of the event. To the best of our knowledge, such problem has never been tackled in the literature.

3. User behavior representation in the context of crisis events

In order to consistently model microblog users with their realistic behavior during events, we propose a user behavior characterization approach that alleviates the shortcomings stated in Sections 1 and 2. An analyzed event is divided into different phases according to its nature/context. This section firstly describes how